Do Actions Speak Louder Than Words? Predicting Influence in Twitter using Language and Action Features

Fatima Al-Raisi, Shadab Alam, Bruno Vavala, Mao Sheng Liu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
Email: {fraisi,shadaba,bvavala,maoshen}@andrew.cmu.edu

***Abstract*—This work explores the connection between language, personality, and influence in a social media network. It clusters users based on two types of features: account activity features and stream content (word) features and compares the usefulness of these different types of features in categorizing users according to their influence and leadership potential in the network. Results of clustering using different sets of features are examined to answer questions about distribution of Twitter users from the influence perspective. These results are compared against distributions of personality traits obtained from previous research on personality types and established assessment tools that measure leadership aptitude and style. Experiments with different clustering algorithms are described and their performance and cluster outputs are reported.**

I. Introduction

This work pursues the research question of how language, personality, and influence are connected. We use Twitter data to analyze users from the perspective of influence. For the purpose of this analysis, we use both user account data and content data to cluster users in different categories according to their leadership or influence abilities. We refer to features extracted from each type of data above as "activity" (or account) features and "word" (or language) features, respectively. We also compare the different types of features in their usefulness for modeling and predicting influence. The underlying reference model is DISC [1], [2]: a behavioral assessment tool that measures different personality traits including leadership aptitude and style. It was first proposed by psychologist William Marston in 1928 and later developed by industrial psychologist Walter Clarke. DISC is a widely used personality assessment tool that is based on studies in various fields and involves surveying a large number of people from different backgrounds, professions, and personalities [3], [4]. A significant portion of this assessment tool is based on the use of language; hence, motivating its use for extracting features to identify influence in tweets. We use this specific tool as a reference for identifying different personality traits associated with DISC categories and focus on the categories of leaders and influential people.

This kind of profiling has many important applications. It can be used to identify points of influence in a large social network. This is of especial interest in places where social media is used as an alternative means to exercise influence or express opinions otherwise not represented in main stream media. For commercial purposes, companies may need to identify influential users to provide them a product or a service so they may recommend it to their followers. This is essentially linked to the "activity shaping" problem in social networks. This kind of analysis can also be used to answer important questions about social media networks such as the similarity in behavioral distribution to patterns/distributions found in larger populations. In addition to answering questions about social media, this kind of analysis can help understand and better model the dynamics of influence, trust, and information propagation. Other applications include targeted advertisement and personalized interface design.

This project is on one hand exploratory work aimed at examining the nature of Twitter data in terms of whether typical DISC distribution patterns can be found in Twitter and whether content/word features are as useful in predicting influence as account/activity features. On the other hand, this work can be viewed as a first step towards automating the task of DISC profiling in social media networks.

II. Background and Related Work

In recent years, a line of research connecting natural language processing and social media analysis has emerged. Several related studies focused on various aspects of personality and interaction including prediction of social relationships and tie strength [5], [6], prediction of Big Five personality traits [7], and prediction of anti-social traits [8]. *Influence* which is an important aspect of personality has been studied using language features or account activity features but has not been explored, to our knowledge, in social media analysis using and contrasting both types of features.

While previous work focused on personality models based mostly on the Big Five personality traits [7], [5], we use DISC model in this work to explore the relationship between language and influence. DISC is an assessment tool that has been developed for different personality analysis purposes including testing leadership aptitude and

style [1], [2]. It basically distributes the population in a space of two dimensions that roughly correspond to 1) people-oriented vs. task-oriented and 2) outgoing/active/fast-paced vs. reserved/reflective/moderately-paced as shown in Figure 1 on page 3. Different variants of the test focus on different aspects of the classification depending on which dimension is more relevant to the problem and its domain. According to DISC studies focusing on leadership aptitude, only 4% of the population falls in the two extremes of leadership aptitude and at most 2% are natural leaders who have strong leadership qualities regardless of training and environment. The majority of the population is found in between not deviating much from the mean in a distribution that resembles a bell curve. The DISC model has been used in different studies for different purposes such as improving team performance through behavioral assessment profiling [9], identifying behavioral factors of individuals in high managerial ranks [3], and even studying the influence of personality style on performance of students in educational settings [4]. In this work we focus on using DISC to answer questions about categorization of Twitter users according to leadership aptitude and style and compare empirical findings to expected distribution based on domain knowledge. We also explore whether language features are as useful as action features (e.g.,#followers, #following,#retweets, etc.) in modeling and predicting influence. We first discover clusters based on user account features and word features separately and then examine whether we obtain similar or different results. We compare the usefulness of each type of features for coming up with clusters that resemble DISC categories and decide whether "actions speak louder than words," "words speak louder than actions," or whether they convey the same information in this context. This also allows us to see differences between the influence aspect of personality and other aspects that are accurately predictable using linguistic content as shown in previous work [7], [8], [6], [5].

III. Data and Feature Engineering

Next we describe the dataset, the different features computed, the motivation and method for feature selection and numerical scaling of features.

A. Data

In this work, we use a subset of Twitter obtained and published in previous work on social user profiling for inferring home locations [10]. The dataset contains network data for 3 million users (profile/account data) and 147 thousand tweet streams. There are about 78 thousand users for which we have both account data and tweet streams.

B. Feature Engineering

We consider two types of features: account (action) features and language content (word) features. Account features include the number of followers, number of friends (following), ratio of the previous two numbers, number of tweets, retweets, and favorite (liked) tweets. The last two are used as an indication of tweet popularity/impact as tweets that tend to be retweeted and liked frequently have more influence and propagate further. In addition to these features that are almost available with the data and required little computation, we also compute the page rank of a user as another account feature. The page rank of a user A is given by the formula:

$$PageRank(A) = 1 - d + d \sum_{i=1}^{n} \frac{PageRank(i)}{L(i)}$$

where n is the number of A followers, $L(i)$ is the number of i's followers and d is the damping factor[1]. Table 1 lists all account (action) features. Note that these features are not all independent. We experiment with different functions and subsets of features for clustering users.

The other type of features is language/content features. These features are based on a bag of words language model in which words are either grouped or treated as individual features. Several linguistic content categorization systems exist including Linguistic Inquiry Word Count (LIWC) system that is commonly used for personality analysis [8], [12], [5], DISC categories which are based on grouping words according to DISC categories: Dominance, Influence/Inducement, Submission, and Compliance, and finally broad categorization of most frequent words into linguistic categories such as function words, common verbs, and pronouns and semantic categories such as social processes, emotions, and work-related words. Both normalized frequencies and tf-idf were used in different experiments to explore the effect of relative weighting in this clustering task.

1) Extracting Content Features: Since this work is exploratory in nature, different sets of word features were used. In one set, words are grouped into categories and one frequency counter is maintained for each category, another set was formed by splitting words into separate features (frequency counter for each word), a third set was formed by including words describing different categories in the DISC assessment, and another variant of the word features was based not on counts but tf-idf scores. The set in which words were split as individual features resulted in very sparse representation of some features so we used the grouped version of the word features (as done in linguistic analysis of most related work). The tweet stream was preprocessed before computing the features. The text was converted to lower case, irrelevant punctuation and other markers were removed, and constant keywords were ignored. However, we did not perform stemming on the tweet stream; weighing the computational cost and information gain we decided that counting variants of the surface word sharing the same stem was not computationally expensive and often important to differentiate.

2) Extracting and Scaling Account/Activity Features: Most action features were readily available in the data. User profiles include the number of followers, friends, retweets, and favorite (liked) tweets. However, in addition to normalizing these

[1]We use the commonly assumed value of 0.85 [11].

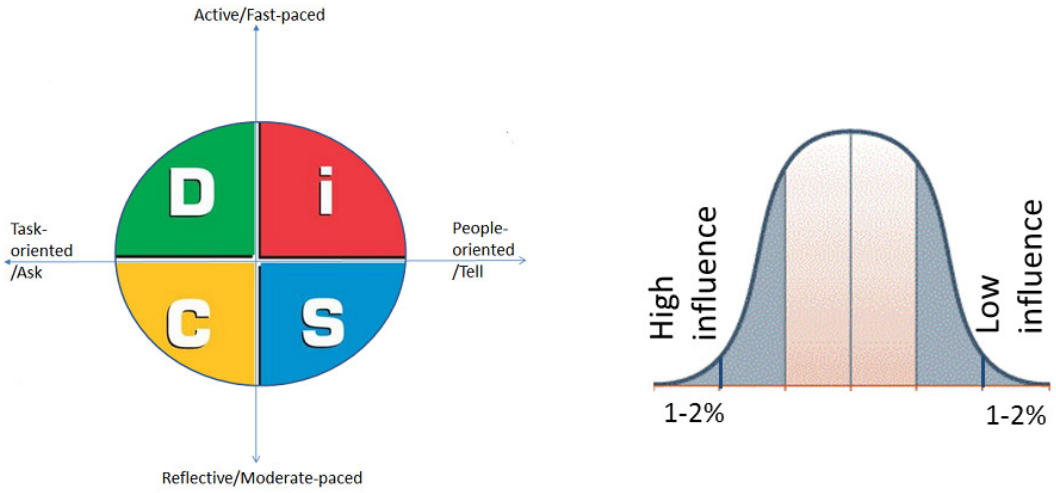

Fig. 1. DISC categories and distribution

activity features	word features
# followers A	LIWC
# following B	DISC
ratio A/B	LIWC + DISC
# tweets	tf-idf of LIWC
# retweets	tf-idf of DISC
# hashtags	tf-idf of LIWC + DISC
# likes	individual words variant of above
pageRank	grouped words variant of above

TABLE I
ACCOUNT ACTIVITY FEATURES AND CONTENT/WORD FEATURES

counts and creating features based on simple functions of these counts (ratios for example), we computed a "pageRank" feature which captures deeper influence in the network than simply the number of followers or the ratio between followers and friends. Interestingly, results show that the pageRank follows a power-law distribution; i.e., there are very few users with high pageRank and there is a long tail of users with small pageRank, while the number of followers have a smoother distribution.

Analyzing user account data, we noticed a wide range of values with a long tail for several key features. This called for scaling of features before feeding the values into clustering algorithms, otherwise the algorithms may yield unexpected results or convergence behavior due to the skewed distribution. The following account statistics are obtained from three million user accounts (3123283 Twitter user profiles). Table 2 shows that the data varies in a wide range with a very long tail across all features. This long tail phenomenon has to be addressed before clustering. Therefore, all features are scaled by the median of the feature set except the *likes* count feature (because its median is 0) which is scaled by 10 times the mean of that count.

3) Feature Selection: Since the number of combined features is very large, Principle Component Analysis (PCA) was done to produce a lower number of linearly uncorrelated features that are most informative. The total number of features exceeds 500, since not all features may be informative and a large number of features may slow the convergence of clustering, we used PCA to represent these hundreds of features in a few eigen components not exceeding 15. Indeed, dimensionality is reduced with PCA and clustering was performed on the projected feature space produced by PCA.

IV. METHOD

Section 2 detailed the different feature types and different linguistic categorizations for word features (LIWC, DISC, individual, grouped). This section presents the clustering algorithms and different experimental settings created from various combinations of feature types and clustering algorithms.

A. Empirical Support for Hypothesis

One of the main questions in this work is whether language and influence are related. We describe an experiment conducted to test the hypothesis that language and influence are related before running clustering algorithms. In this experiment, users were ranked according to each action feature (#followers, #following, etc.), resulting in n different rankings/lists (where n = #action features), then the top 5% of users in each list were extracted, the pair-wise intersection of user lists (i.e., intersection of top 5% users according to each pair of features) was obtained, and finally the union of resulting sets was taken.

Feature	min	max	mean	median	sum
# FRIENDS	0	695509.0	719.4	218	2246906359.0
# FOLLOWERS	0	11060753.0	1348.48	136.0	4211702993
# TWEETS	0	982934.0	2319.71	272.0	7245121544.0
# LIKES	0	3200.0	27.41	0.0	85628538.0

TABLE II
ACCOUNT DATA STATISTICS

The final set contained the top 5% users according to action features. Initially, the experiment was designed to simply take the intersection of all n lists and regard that intersection as the top 5% influential users according to action features but that intersection was almost empty[2] which suggests that these features were not redundant and that each feature targets a different "action" and therefore possibly different kinds of users. The other option was to simply union all top 5% users obtained from rankings by different features but that set may contain users that are not as influential as those who are ranked highly by more than one feature. Taking the pair-wise intersection and then taking the union of all resulting sets is a balanced option in between. The 5% lowest rank users were sampled following a similar procedure.
We then examined the language use for these two sets of users that are on different extremes according to action features. A clearly different use of language across all linguistic categories is observed. The variation is measured in differences in (normalized) frequencies of words across categories as shown in Figure 2. One observation is that the top 5% users tend to use more words (i.e., express themselves more) and that the ratio between the two sets of users varies across categories from double to more by a third or less. This supports the intuition that language and influence are related. The following section describes the clustering experiments conducted to further examine this hypothesis and answer other questions about influence distribution in Twitter.

B. Clustering

Clustering was done using three different clustering algorithms: k-means, EM, and spectral clustering. In each experiment users were clustered according to word features and account features, separately. The resulting sets of clusters are examined for similarity and overlap. The idea is that if we cluster using action features and cluster using word features separately and then find that the resulting clusters are similar and overlap then we can infer that these different sets of features (actions and words) model the same phenomenon: influence, and that although they are different in nature they are strongly related as they can make similar predictions about the same phenomenon. Algorithm 1 on page 4 is high-level description of the clustering and analysis steps.

k-means: We experimented with different values of k. Based on the problem domain, however, we selected 4 as it reflects the number of main categories in DISC. Although k-means is suitable in settings where the data is expected to be separable, we noticed that clusters we obtained are dense around the mean with far fewer points spread further. So even with k-means we were able to obtain clusters that were clearly distinct.

[2]The intersection of all top 5% lists included only 6 users from the original list of 78 thousand users.

Algorithm 1 Cluster and Analyze

```
for each feature set S do
  for each clustering algorithm A do
    cluster users according to S using A
  end for
end for
for each clustering algorithm A do
  examine overlap between action-based clusters and word-based clusters
  examine similarity of clusters obtained using different linguistic content categorization
  examine relative sizes of clusters in each clustering
end for
return overlapping clusters, cluster size distribution, and corresponding algorithm A
```

EM: Since EM is suitable for soft clustering where clusters are expected to overlap, we clustered the users using EM on mixture-of-Gaussian models. This was motivated by the large number of data points ($>$ 78K users) calling for the applicability of the central limit theorem as a reasonable assumption and more significantly that the overall distribution of influence resembles a normal distribution as depicted in II on page 3.

Spectral: We also applied spectral clustering to see if it confirms results of other clustering algorithms or behaves differently. Spectral clustering is further motivated by its applicability in settings where clusters may overlap. We noticed that spectral clustering does not scale with a large dataset. To successfully apply spectral clustering, k-means was first run to reduce the dimensionality of the data and then spectral clustering was run on the dimension-reduced dataset. Different values of k ranging from 700 to 50 were tried, spectral clustering scaled only to the smallest dataset; i.e., the maximally reduced set with 50 means.
In the following section, we present results and compare the performance of clustering algorithms.

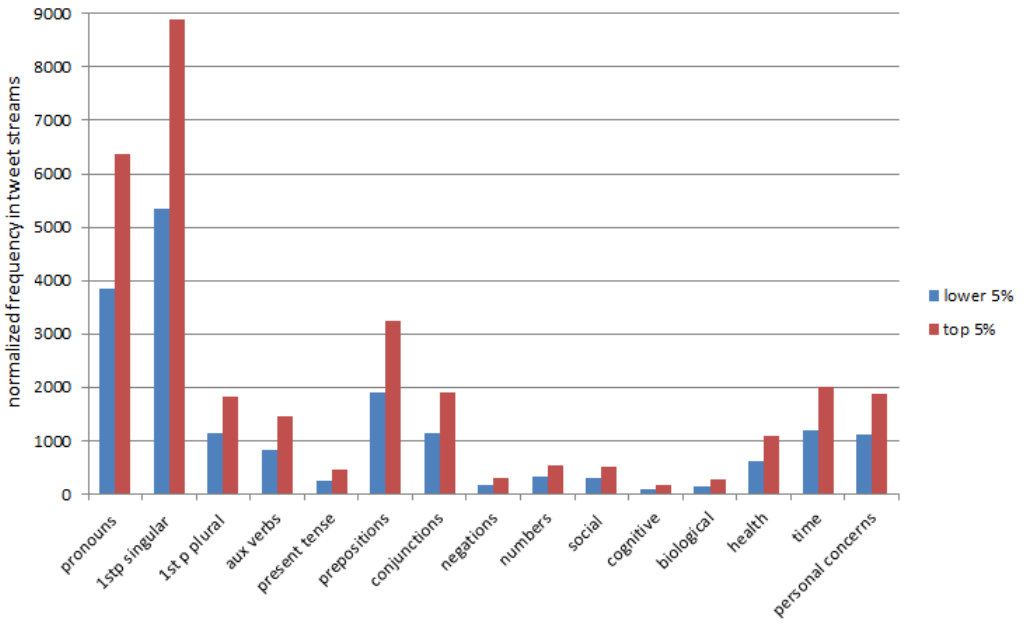

Fig. 2. language use in different groups of users

V. Results

A. Comparison of Different Clustering Algorithms

Table 3 summarizes differences in performance and output of clustering algorithms. The runtime reported is based on running the algorithms on the largest data set of profiles ($>$ 2.5 million users). The runtime for spectral clustering includes the runtime of running k-means with k=50 first and then running spectral clustering on the result.

B. Cluster Overlap Results

Clusters based on action features and those based on word features are examined for overlap. Overlap between two clusters is measured in the number of users they have in common. The idea is that if action-based clustering and word-based clustering result in respective clusters that overlap; i.e., have many users in common, we conclude that action features and word features can mirror each other in making similar predictions and therefore are related. Figure V-B shows overlap patterns for clustering using different sets of features. A diagonal means that overlap is observed in all four clusters. A partial diagonal means that overlap is observed only in some subset of clusters. We present the plots that represent common patterns across algorithms and discuss these findings in the analysis section.

VI. Analysis and Discussion of Results

The following are observations from the experimental results along with analysis for each result.

1) Clustering based on account features and on word features show that the sizes of clusters produced closely match the typical size of population categories according to DISC with two small clusters representing the extremes on influence abilities and two large (possibly overlapping) clusters representing most of the population. In fact, all clustering combinations almost always produce a skewed distribution of cluster sizes which matches the typical distribution of individuals according to the theory of influence and leadership. It is also observed that the distribution is more skewed when word features are used to cluster. We've attempted further analysis on the word content of these clusters in order to draw reliable conclusions about the nature of these clusters. This is discussed in cluster identification section below.
2) Similarity was found between clusters obtained using word features from different linguistic content categorization systems. More specifically, adding DISC word categories to LIWC does not significantly change clusters obtained from LIWC word categories. This shows that the latter is a comprehensive categorization that subsumes DISC categories. However, clustering based on DISC categorization of linguistic content results in better alignment with clusters obtained from action features. This agrees with the claim that DISC is designed to specifically target the influence aspect of personality.
3) Similar clusters are obtained from account features in original spaces and PCA projection space. This result is not surprising due to the small number of action features that even when projected using PCA result in principle components that are very similar to the original representation.
4) More importantly, a clear overlap is observed between clusters based on action features and clusters based on word features.

	run-time	scalability with		cluster sizes
		# features	# data points	
4-means	1m 1s	✓	✓	12,19,27,42(%)
EM	5m 3s	✓	✓	6,13,36,45(%)
Spectral	42m 44s	✗	✗	6,18,35,41(%)

TABLE III
COMPARISON OF DIFFERENT CLUSTERING ALGORITHMS

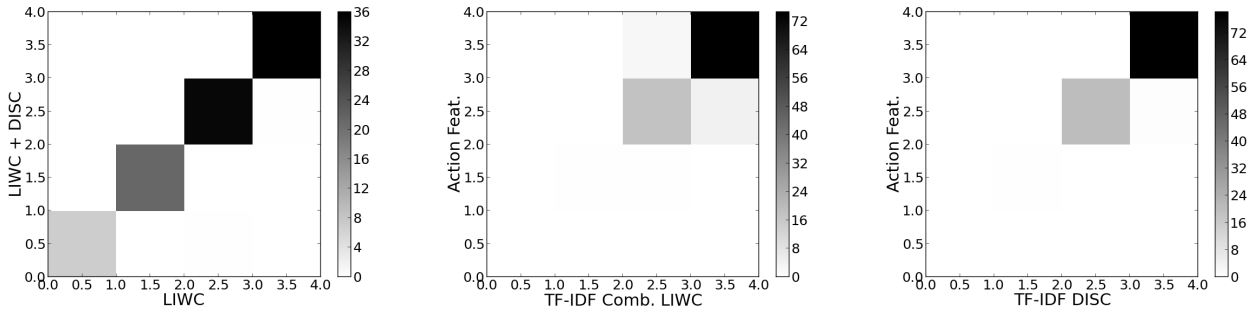

Fig. 3. Overlap between clusters obtained using different feature sets

5) Different clustering algorithms produce different clusters for both action and word features yet EM and spectral clustering results are closer to each other. This shows that each algorithm handles the data differently leading to different clusters. However, the overall relative sizes of clusters and their densities are very similar. We suspect that each of these algorithms is best at modeling a specific aspect of the intricate phenomenon of influence for example: degree of influence vs. style of influence.

VII. CLUSTER IDENTIFICATION

The objective of cluster identification is to label instances in clusters according to influence categories. In the initial experiments, our analysis was not limited to the sizes of clusters. We examined the word content for clusters based on user account (activity) features. The analysis was to see if a mapping exists between clusters obtained from account activity features and those obtained from word features. Although the use of language shows commonalities among clusters, there are variations in the word patterns from cluster to cluster suggesting some differences in the use of language across clusters obtained from account/action features only. There are at least two clusters that seem more similar and two that seem to exhibit different use of words. This agrees with the distribution of cluster sizes with two small clusters (different use of language) and two overlapping clusters (similar use of language).

We also checked membership of the top 5% users sampled according to action features to see if they belong to the same word-based cluster which we can use to identify the cluster of most influential users. These users were distributed across all clusters.

We sampled tweet streams of users from each cluster (closest to the centroids in k-means) and examined the content of

Fig. 4. visualization of language content of sample streams: small cluster vs. large cluster

the small vs. large clusters that we expect are the influential and ordinary clusters, respectively. Indeed, the sample from the small cluster streams shows that the discussions and topics were about work, strategy, marketing, and topics that are in general very different from those found in the other sample which are mostly related to personal concerns and daily activities as depicted in Figure 4 on page 6. This shows that the small clusters, as expected from background knowledge on typical influence distribution, correspond to influence extremes. The limitations of these samples should be taken into consideration, however, as these samples are small and streams are noisy.

A. Challenges and Remarks

Data Representation and Sparsity: Coming up with useful informative features require revisiting domain specific algorithms and implementing them. For example, pageRank features. Four different versions of word features and different grouping combinations and linguistic models were used. We tried running the clustering algorithms using all combinations

but some resulted in sparse representation of the data so PCA and clustering algorithms could not be run. Although this was useful for feature selection (eliminate features leading to sparsity), it was counter-intuitive as some of these features were carefully selected based on domain knowledge (i.e., relevance to DISC).

Dimensionality: With the large number of user account and stream content features, both feature extraction and data clustering require significant computation power and memory. Most of the experiments were run on a dedicated server yet some algorithms had excessively long run times or crashed due to memory errors. We attempted different clustering algorithms including DBSCAN (for density-based clustering) and spectral clustering which was useful in understanding scalability limitations of different clustering algorithms and possible approaches to address these limitations like preprocessing using k-means prior to spectral clustering. Moreover, we attempted to cluster in different spaces: original feature space and feature space from PCA. Running in original feature space required considerably more computational time and power so clustering was mostly done on the transformed feature space.

Unsupervised learning of latent parameters: With no labels and no explicit correspondence between "influence" and language in the noisy data present in social media, the problem of identifying users with leadership potential is a challenging one. Using a latent variable model such as LDA or a combination of semi-supervised and active learning methods may be useful in moving from clustering to classification and realizing the goal of automated leadership profiling in online communities.

VIII. Conclusion

The overlap between clusters obtained from word features and those obtained from action features suggests the usefulness of both types of features in predicting influence. This answers the main research question posed in this work and shows, based on empirical evidence, that language and influence are connected. It is important to contrast results obtained from these two different sources of information especially with the more recent phenomena of bought followers and likes that can render account (non-content) features less reliable. Our experiments show that using different systems for analyzing the language content of Twitter can lead to variations in results and that DISC seems to be the most appropriate tool for studying influence. The influence distribution found in this sample of Twitter resembles that obtained from large population statistics. It is highly skewed and shows two minority categories that are clearly separated and two larger overlapping categories. Manual labeling of sampled instances from different clusters shows different use of language and agrees with theoretical and empirical results suggesting that influential users are a minority.

Appendix: Supplemental Material

k-means is typically used in cases where clusters are expected to be separable. However, the following figures are provided to illustrate that even when applying k-means to cluster users according to language features that are expected to result in some overlaps, we get clusters that are dense around the mean and clearly distinct. Since the data is multidimensional, a special distance function was designed to compute the distance between clusters and present their relative sizes and distances from each other in 2D. The two selected visualizations are for clusters obtained from tf-idf features on LIWC and DISC linguistic categorizations. The graphs show only three clusters because in one case the fourth cluster is too small to be seen relative to the other three clusters and in the other the fourth cluster is at a large distance from the other three clusters making it not possible to fit all four clusters in one reasonably scaled image.

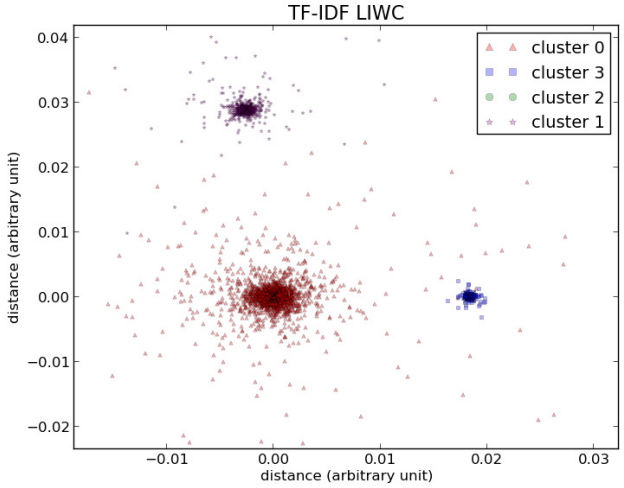

Fig. 5. Visualization of of k-means clusters based on tf-idf LIWC word features

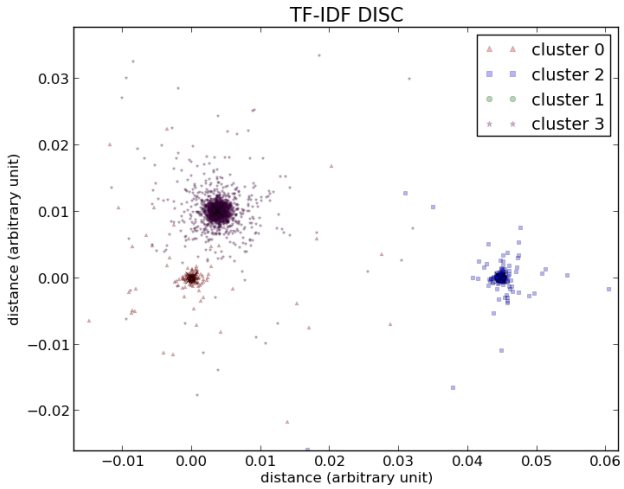

Fig. 6. Visualization of k-means clusters based on tf-idf DISC word features

References

[1] "DISC Instrument Validation Manual," http://www.coachannette.com/pdfs/DISCValidityManual.pdf, 2004.

[2] "DISC Classic®Validation Report," https://www.inscape-exchange.com/downloads/marketing_support/researchreports/DiSCClassicValidationResearchReport.pdf.

[3] G. Beamish, "How chief executives learn and what behaviour factors distinguish them from other people," in *Industrial and Commercial Training*. Emerald Group Publishing Limited, 2005, vol. 37, pp. 138–144.

[4] P. Blignaut and A. Naude, "The influence of temperament style on a student's choice of and performance in a computer programming course," *Comput. Hum. Behav.*, vol. 24, no. 3, pp. 1010–1020, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2007.03.005

[5] S. Adali, F. Sisenda, and M. M. Ismail, "Actions speak as loud as words: predicting relationships from social behavior data," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, New York, NY, USA, 2012, pp. 689–698. [Online]. Available: http://dx.doi.org/10.1145/2187836.2187930

[6] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 211–220. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518736

[7] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 149–156. [Online]. Available: http://dx.doi.org/10.1109/passat/socialcom.2011.33

[8] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets." in *ICMLA (2)*. IEEE, 2012, pp. 386–393. [Online]. Available: http://dblp.uni-trier.de/db/conf/icmla/icmla2012-2.html#SumnerBBP12

[9] J. Duck, "Making the connection: Improving virtual team performance through behavioral assessment profiling and behavioral cues," in *Developments in Business Simulation and Experiential Learning*, vol. 33, 2006, pp. 358–359.

[10] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations." in *KDD*, Q. Yang, D. Agarwal, and J. Pei, Eds. ACM, 2012, pp. 1023–1031. [Online]. Available: http://dblp.uni-trier.de/db/conf/kdd/kdd2012.html#LiWDWC12

[11] A. Altman and M. Tennenholtz, "Ranking systems: The pagerank axioms," in *Proceedings of the 6th ACM Conference on Electronic Commerce*, ser. EC '05. New York, NY, USA: ACM, 2005, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/1064009.1064010

[12] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, 2010. [Online]. Available: http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html