Semantics and Homothetic Clustering of Hafez Poetry

Arya Rahgozar **Diana Inkpen**
School of Engineering and Computer Science
University of Ottawa, Canada
800 King Edward Ave. Ottawa, ON, K1N 6N5
arahgoza@uottawa.ca diana@site.uottawa.ca

Abstract

We have created two sets of labels for Hafez[1] (1315-1390) poems, using unsupervised learning. Our labels are the only semantic clustering alternative to the previously existing, hand-labeled, gold-standard classification of Hafez poems, to be used for literary research. We have cross-referenced, measured and analyzed the agreements of our clustering labels with Houman's chronological classes. Our features are based on topic modeling and word embeddings. We also introduced a similarity of similarities' features, we called homothetic clustering approach that proved effective, in case of Hafez's small corpus of ghazals[2]. Although all our experiments showed different clusters when compared with Houman's classes, we think they were valid in their own right to have provided further insights, and have proved useful as a contrasting alternative to Houman's classes. Our homothetic clusterer and its feature design and engineering framework can be used for further semantic analysis of Hafez's poetry and other similar literary research.

1 Introduction

Chronological classification of Hafez poetry was done by Houman, in his book (Houman, 1938). He partly hand-classified Hafez's poems in 1938, based on the semantic attributes engraved and encrypted in the ghazals. Houman's labeling has been the gold-standard of chronological classification for Hafez, and Rahgozar and Inkpen (2016b) used them as training data for supervised learning to predict the rest of the ghazals. We used similar semantic features, but instead we conducted unsupervised learning (clustering experiments) to create alternative labels to those of Houman.

Houman's classification was based on the premise that artist's mindset and worldview changed throughout his lifetime and this change was reflected in his art, in this case, poetry. Hypothesising about the evolutionary reflection of this chronological worldview in the semantics of Hafez's art and capturing it, was Houman's intension; so was ours, but by using machine learning. For example, Houman believed that the old Hafez was more introverted than the young. Houman explained in detail that these worldview characteristics and their interpretations were buried in the semantic attributes of Hafez's highly indirect, multi-layered and equivocal ghazals, intertwined among couplets' and hemistiches' surface meaning, but differently throughout his life.

1.1 Problem Statement

We hope that the chronological classification of Hafez would facilitate interpretations and demystify the depth of meaning in his majestic poetry. In this work, we used clustering as a semantic analysis tool to assist with literary investigations of Hafez's poetry. As a result, we have produced new unsupervised labeling standards for Hafez corpus[3]. We have also conducted what we refer to as *homothetic clustering* experiments, using similarity transformations as features, discussed in Section 2.5. We have performed semantic analysis, partly discussed in Section 4, using a topic-modelling visualization interactive tool.

Although the fundamental question was to find out how consistent our semantic-based clustering would be with Houman's chronological classification, and to establish a verification experiment

[1] Persian philosopher and poet.

[2] Popular form of Persian poetry with specific rhyme and rhythm, consisting of about ten, seemingly independent couplets; Ghazal is interchangably used with the word poem here.

[3] Our Hafez corpus will be available, alternative sources for Hafez corpus are https://ganjoor.net/hafez/, http://www.nosokhan.com/ and https://www.hafizonlove.com/

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 82–90
Minneapolis, MN, USA, June 7, 2019.

against Houman's labeling, we set to achieve the following objectives:

- Semantic Feature Engineering;
- K-Means Clustering: Automatic Semantic Labeling;
- Similarity Feature Transformation as Homothetic Clustering;
- Multi-label Semantic Analysis and Visualization: Houman's, plus Machine Labeling.

We also wanted to see if homothetic features could qualify our unsupervised method as a guided or quasi-semi-supervised labeling.

2 Methodology

Our focus was to observe the performance and identify the semantic features that provided us with the best clustering results, measured by *Silhouette*. We were also interested to find out which features produced more consistent results with Houman labels. To measure interagreements we used *kappa* and other measures. In all the experiments, the clustering algorithm was K-Means to focus on the effects of features.

2.1 Corpus Work

Our bilingual[4] Hafez corpus had six chronological classes labeled by Dr. Houman[5] that were logically enumerated from *Youth* to *Senectitude*, therefore they could be logically consolidated into valid three classes, while maintaining their sequential order. Houman only labeled 248 poems out of 460 total confirmed Hafez ghazals, and we only considered those poems for clustering, so that we could cross-reference, verify and compare their Houman-classifications with our clustering generated labels or classes.
We applied the *white-space*[6] character and zero-width joiner (ZWJ), wherever it was needed in our corpus, so that the linguistic properties of Persian words and their inflections were maintained consistently.

[4]Persian-English

[5]Dr. Houman labeled Hafez in about 1317 SH (1939 AD).

[6]Persian words can be multi-words; white-space is a transparent character linking the sub-tokens, for example **daneʃ âmuz** means student, is one word, but is written as two.

2.2 Preprocessing

We followed (Asgari and Chappelier, 2013) for our preprocessing steps:

- Tokenization
- Normalization
- Lemmatization
- Filtering

In our preprocessing we removed the stop-words and the tokens that occured only once. We built the dictionary of documents, every document being a poem (ghazal). Then using the bag-of-words, we set up and transformed the corpus into vector representations. We built the TF-IDF[7] vectors accordingly. We initialized LSI, LDA[8], Log-Entropy (Lee et al., 2005) and Doc2Vec (Le and Mikolov, 2014) objects using both the Persian and Persian-English corpus as training. We used gensim library (Řehůřek and Sojka, 2010) and used HAZM[9] Python library for Persian pre-processing tasks, such as *lemmatization*.

2.3 Clustering Evaluation Indices

We followed metrics and clustering agreement techniques and scores[10] to measure our performance results in comparison with Houman's chronological labels. A value of one indicated perfect consistency.

- *Inertia*: Within-cluster sum of squared criterion, which K-Means clustering tries to minimize; the lower the inertia is the better.
- *Homogeneity*: Average single Houman class poems' distance to the center of the clusters; clusters are homogeneous if they only contain poems of a single Houman-class;
- *Completeness*: A measure of parallel correspondence between Houman classes and our clusters;

[7]Term frequency/inverse document frequency is a measure of term's importance among documents in the corpus.

[8]A high number of topics were pointless given our small corpus size, but we chose ($5 < Topics - Number < 20$), based on Silhouette convergence, in each experiment setting.

[9]https://pypi.org/project/hazm/

[10]http://scikit-learn.org/

- *V Measure*: Homogeneity = HOM, Completeness = COM:

$$2 * (HOM * COM)/(HOM + COM)$$

- *Adjusted Random Index* (ARI): Is a similarity measure between clusters by pairwise comparisons of cluster and Houman class poems, E = Expected:

$$ARI = (RI - E(RI))/(max(RI) - E(RI))$$

- *Adjusted Mutual Info*: Is a symmetric measure of dependence between our cluster membership and the Houman-class:

$$\frac{MI(U,V) - E(MI(U,V))}{max(H(U),H(V)) - E(MI(U,V))}$$

- *Silhouette*: Is a measure of cohesion and distinctive quality to separate clusters, that is the mean of a and b, $(b - a)/max(a, b)$, where a and b are aggregated intra-cluster and nearest-cluster distances of each poem.

- *Cohen's kappa* measures the consistencies between two sets of labels, generated by classification or clustering[11]:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

2.4 Feature Engineering

The variant of TFIDF we used was based on a logarithmically scaled frequencies of term i in document j in a corpus of D documents:
$weight_{i,j} = frequency_{i,j} * log_2 \frac{D}{document - freq_i}$
The LDA[12] implementation followed (Hoffman et al., 2010); base code was found here[13]. We kept the default parameters when initialized the LDA model, except setting wokers equal to 8. For the LDA driven similarities, we only set the number of topics and passes to 5.
Doc2Vec[14] implementation followed (Mikolov et al., 2013). We set the parameters as follows: vector size=249, window=8, min count=5, workers=8, dm = 1, alpha=0.025, min alpha=0.001, start alpha=0.01, infer epoch=1000.

[11] en.wikipedia.org
[12] https://radimrehurek.com/gensim/models/ldamulticore.html
[13] https://github.com/blei-lab/onlineldavb
[14] https://radimrehurek.com/gensim/models/doc2vec.html

2.5 Homothetic Features: Sim^2

Homothetic transformations are frequently used in transferring arguments amongst economic models. Intuitively, one could think of the concept as similarity of similarities. In our case, for every poem in the corpus, represented as LDA-driven vector, we derived and formed a new vector, consisting of calculated *Cosine* similarities or distances from that poem to a subset of hand-picked poems, we refer to as anchors. Anchors were chosen for semantic reasons to guide the clustering towards Houman's classes. By these similarity measures to the anchors, we formed a new vectorized corpus. In other words, we used *Cosine* similarity as a transformation function from one vector space to another, before we measured their Euclidean distances, in a clustering procedure such as K-Means.

Data: Hafez Corpus
Result: Generate labels
read corpus and anchor instances;
tokenize, remove stop-words and tokens-once;
normalize, lemmatize;
create *bag-of-words, TF-IDF*;
initialization LDA;
create *LDA-driven* similarity index;
while *not at end of the corpus* **do**
  while *not at end of the anchors* **do**
    calculate *similarity* Measure;
    append to vector list;
    go to the next anchor;
  end
  write document similarities: *Sim-Corpus*;
  go to the next document;
end
set k clusters;
cluster (*Sim-Corpus*);
produce predictions;
Algorithm 1: Homothetic Clustering, Sim^2

2.5.1 Homothetic Properties

Similarity transformations are not necessarily linear, as we ran into the equality contradiction of summation of two square roots of polynomials and that of one, which proves the nonlinearity property, in a 3D Euclidean space:

$$f(u) + f(v) \neq f(u + v)$$

Similarity transformations also maintain *homothetic* properties, a monotonic transformation of a

Feature	Inertia	Homog.	Comp.	v-meas.	ARI	AMI
LogEntropy	238	0.017	0.015	0.016	-0.004	0.008
LSI	237	0.004	0.004	0.004	-0.003	-0.004
LDA-TFIDF	233	0.003	0.009	0.005	0.013	-0.007
LDA	233	0.006	**0.023**	0.009	-0.007	-0.004
Doc2Vec-P	1445	0.010	0.010	0.010	-0.008	-0.002
Doc2Vec-PE	338	0.020	0.017	0.018	0.018	**0.010**

Table 1: K-Means Performance, $(k = cls = 3)$
cls = number of classes

homogenous function for which the level sets were radial expansions of one another. In Euclidean geometry, a homothety of factor k magnifies or *dilates* distances between points by $|k|$ times, in the target vector-space. Risk of overfitting and its divergence was also empirically suspected to be higher and quicker. The properties of Homothetic functions were proven by (Simon and Blume, 1994):

$$v(tx) = g(u(tx))$$

$$g(t^k u(x)) = g(t^k u(y)) = g(u(ty)) = v(ty)$$

We have demonstrated empirically, that the homothetic clustering procedure we used here, was effective to increase Silhouette score and showed tractable interpretations, when used against our small poetry corpus of Hafez. The average complexity of the homothetic clustering was the same as the complexity of the clustering method it uses. In this case, we used K-Means with polynomial smoothed running time, therefore the complexity was the number of samples n, times the number of iterations i, times the number of clusters k:

$$Complexity(Sim^2) = O(n.i.k)$$

3 Experiments

In the first set of experiments, we used different semantic features for clustering. We then passed the vector representation of the labeled portion of the corpus to K-Means[15] for clustering ($k = 3, 6$). Then we compared the clustering labels with Houman labels. The Table 1 shows the results. As we see, the Doc2VecPE feature ranked at the top in *Homogeneity*, *V-measure*, *ARI* and *AMI*. The LDA feature obtained the best in *Completeness* compared to other features. As we see in Table 2 The pure Persian *Embedding, (Doc2Vec-P)* showed the highest *Silhouette*[16], while adding English[17] to the corpus brought this measure a bit lower and still maintained second rank compared to all other features.

[15] http://scikit-learn.org/

[16] Defined in Section 2.3

[17] English translation of the poems by Shahriari, were inline with the Persian version, when the translation was available.

Feature	3cls-Silhouette	6cls-Silhouette
LogEntropy	0.001	-0.000
LSI	0.001	-0.002
LDA-TFIDF	0.037	0.097
LDA	0.059	0.109
Doc2Vec-P	**0.560**	**0.528**
Doc2Vec-PE	0.530	0.471

Table 2: K-Means Performance
P=Persian, E=English

Feature	Inertia	Homog.	Comp.	v-meas.	ARI	AMI
HRP	0	0.034	0.035	0.034	-0.001	0.004
HEP	0	0.024	0.024	0.024	-0.006	-0.006
RND	0	0.021	0.022	0.021	0.001	-0.009

Table 3: Sim^2 Performance
$(k = anchors = cls = 6)$

3.1 Homothetic Clustering Experiments

Houman (1938) picked a representative poem for each of his classes. For every poem of the labeled portion of the corpus, we calculated the LDA-based similarities to either three (or six) anchor poems, depending on the intended clusters. The resulting vector-space had three (or six) dimensions. We called this Houman Representative Picks (HRP). In a separate set of experiments, we also picked six poems as anchors, three poems from either extreme peripheries of the Houman's labeled poem classes, that is three from the earliest *Youth* class, and three from the latest period ranked in the *Senectitude*. We referred to this experiment's feature set, Houman Extremal Picks (HEP). Or in case of the three classes HEP, we picked two extremal poems and one from central poem from class two, mid-age. RND stands for random picks. We always maintained that the number of anchors matched with the number of intended clusters: $(anchors = k = 3, 6)$, shown in the tables.

As we see in Table 3, HEP, HRP and RND maintain zero *Inertia*, which is an indication of perfect inner cohesion of the clusters. HRP has about 3% as the highest *Homogeneity*, which was higher than that of the challenger, Table 1. LDA had the highest *completeness* as challenger, while Doc2Vec-PE had the highest *AMI*. Both HRP and HEP champion models with similarity features also entailed higher *Silhouette* scores in clustering (Table 4) than the one achieved by

Feature	6cls-Sil.	6cls-Kap.	3cls-Sil.	3cls-Kap.
HEP	0.837	0.004	0.695	-0.014
HRP	0.903	**0.034**	0.824	-0.006
RND	**0.945**	-0.052	0.821	-0.001

Table 4: Sim^2 Performance, (kappa with Houman)

the challenger model, with word-embedding features. Only HRP showed slight resemblance with Houman's classes, as kappa indicated in the same Table. This means that Houman's poems that he mentioned in his book as their class representatives, while explaining his methodology, had a better homothetic guiding power than the actual extremal poems of his classified corpus, when we used them as anchors.

The number of LDA topics in multiple K-Means runs, affected the Silhouette score, but mostly converged in around 5 to 15 topics, depending on the feature set. To avoid local-optima, it was also important to iterate through K-Means algorithm enough times to attain an optimum Silhouette score while targeting the right number of LDA topics, to achieve the best possible clustering quality. Our Homothetic experiments achieved best *Silhouette* scores with 6 LDA topics. In all homothetic and non-homothetic clustering experiments, number of clusters $k = 6$ and $k = 3$, achieved the highest silhouette scores, in their experiments group respectively, $k = anchors$. In homothetic experiments, $k = 6$ clusters always produced both better kappa[18] and silhouette, regardless of the number of anchors being 3 or 6.

We also compared the consistency of HEP Sim^2 clusterer with the challenger (Doc2VecP) model. The Spearman correlation was 0.86. Noteworthy, the Cohen's linear and nonlinear *Kappa* were 0.58 and 0.43 respectively, between these two independent clusterers.

Our Student's t-test did not support the claim that anchors guided the Sim^2 clustering to have a significant consistency with Houman classifications, when we compared the effects of HEP and HRP anchors with randomly selected 6 anchors instead, using *kappa*. Although random anchors were selected with the proviso that they came from different Houman classes.The *Silhouette* of Sim^2 clusterer with random anchors was close to that of HEP and HRP, very high.

[18]Comparing only when $k = cls$.

Duplicity, Sufi and Abstemious	A	B	C
Doc2Vec-P	56, 19, 22	12, 2, 3	17, 3, 4
HRP	31, 11, 13	30, 5, 6	24, 8, 6
HCEP	19, 4, 5	53, 15, 13	13, 5, 7
Vision, Barmaid, Knave	B	A	C
Doc2Vec-P	18, 11, 17	58, 39, 67	8, 10, 0
HRP	17, 19, 19	29, 26, 29	38, 15, 0
HCEP	51, 38, 63	18, 11, 14	15, 11, 0
Expedient, Guru, Pub	C	A	B
Doc2Vec-P	1, 9, 0	6, 44, 1	2, 11, 0
HRP	4, 22, 5	1, 21, 0	4, 21, 1
HCEP	3,14, 1	0, 14, 0	6, 36, 0

Figure 1: Tracing Clusters of Terms

4 Analysis and Discussion

We used the Persian part of the corpus for this section, suffices to demonstrate the semantic values of our new sets of labels.

4.1 Cycle of Words

More rigorous analysis should be done by literary scholars, but as a sample of examination, we constructed in Figure 1 as follows. We counted the Houman labeled poems in each cluster and calculated their percentages to decide the highest resemblance of each cluster with its closest Houman class. In case of a tie, we did the same for the other clusters and then tracked back to maximize an overall resemblance. HRP and HEP were constructed as explained in Section 3.1. Then we considered a cluster of terms, relevant to Houman's representative poems and his semantic constructs (Houman, 1938). For Youth class (A), we chose three terms: Duplicity (*rıâ*), Sufi (*sufi*) and Abstemious (*zâhəd*), and for Mid-age class (B), we chose Vision (*nazar*), Barmaid (*sâqi*), Knave (*rənd*) and finally for the Senectitude (C), we chose three representative terms of Expedient (*masləhat*), Guru (*pır*), Pub (*meikade*). Then we counted the frequency of the terms in each cluster, as per the closest Houman-class. Each cell in Figure 1 contains frequencies of three terms respectively.

If we trace any effect of anchors' semantics in the final homothetic clustering result, we observed that HRP had slightly stronger resemblance with the Houman classes as it was also measured by higher homogeneity and completeness in Section 3.1. Both HEP and HRP showed bet-

ter overall balanced distribution in terms of size of each cluster compared to Doc2Vec-P, which was also reflected in the higher silhouette score from Section 3.1. Although both HEP and HRP showed stronger correlation with Houman-classes than Doc2Vec did. HEP was also stronger in discriminating against class A and C which was attributed to its original anchor poems purposely picked from the same peripheries of the chronological Hafez corpus. This simple example, therefore, was consistent with the assumption that similarity measures transferred the information to the clustering and guided it as per the semantics of the *anchored* poems.

4.2 Semantic Analysis

Each poem's new label provided new perspective and insights, to enable us interpret Hafez's poem better, by investigating the semantic characteristics of its associated cluster, in conjunction with its Houman classification. We could visualize the corresponding cluster, using *LDAvis* topic modelling (Sievert and Shirley, 2014) who introduced and used *Relevance* measure. (2012) defined and developed *Saliency* as part of Termite visualization tool.

For example, we selected to analyze a poem, number 230 from the Houman labeled portion of the corpus, which was the number 143 in Ganjour[19]. On the one hand, we saw that this poem belonged to class 5 or *before-senectitude* of Houman's classification. On the other hand, we looked at the top 30 terms of the topic 3 which was central in PCA depiction of 5 LDA topics, Figure 2, which corresponded with our new label 1 cluster poems generated by Sim^2 clusterer. The words *old (pɪr), Heart (dəl), Love (əʃq), Guru (pɪr ə moqân), Sadness (qam), Ocean (dariâ), Circle (dâyərə), Want (talab), Destiny (kâr), Sigh (âh)* were not only semantically consistent between the two classifications, but they also provided us with a tangible context to better understand and associate with the poem.

Interacting with the visualization tool revealed other themes associated with this previously known as *before-senectitude* poem, that for example, showed a topic 2 at the left of PC1 line, having top salient words such as *jewel (laəl), gal (iâr), sun (xorʃɪd), earth (xâk), hand (dast), heart (dəl), joy (xoʃ), laughter (xandân), love (əʃq), flaw (əib)*. This indicated that the traces of material world and its desires still equally existed and decorated Hafez's poetry, even during those mature years of his life, but he perhaps used these words more metaphorically and mystically.

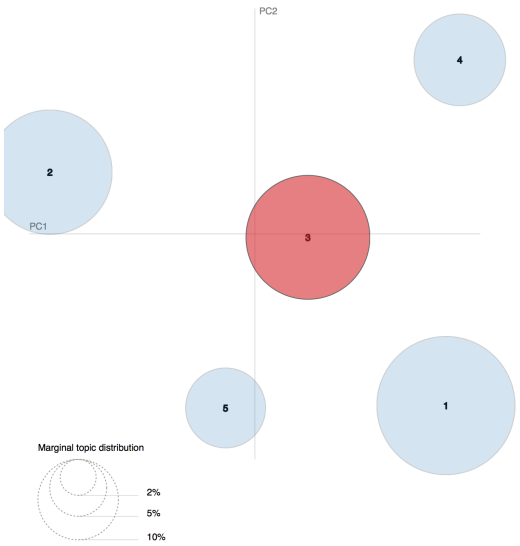

Figure 2: Intertopic Distance Map

For years my heart was in search of the Grail What was inside me it searched for on the trail

That pearl that transcends time and place Sought of divers whom oceans sail

My quest to the Magi my path trace One glance solved the riddles that I Braille

Found him wine in hand and happy face In the mirror of his cup would watch a hundred detail

I asked "when did God give you this Holy Grail?" Said "on the day He hammered the worlds first nail!"

Even the unbeliever had the support of God Though he could not see Gods name would always hail.

All the tricks of the mind would make God seem like fraud Yet the Golden Calf beside Moses rod would just pale.

And the one put on the cross by his race His crime secrets of God would unveil

Anyone who is touched by Gods grace Can do what Christ did without fail.

And what of this curly lock that's my jail Said this is for Hafiz to tell his tale.

5 Related Work

Semi-supervised concepts, prototype and anchors have been discussed in the literature (Zhang et al., 2015), but our approach was new in that no label was directly used in the algorithm. Instead, instance similarities to a few labeled instances formed the entire vector space as their feature set,

[19] https://ganjoor.net/hafez/ghazal/sh143/

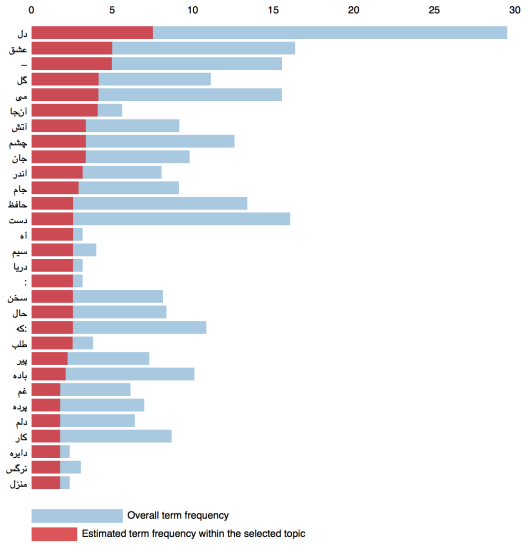

Figure 3: Top 30 Most Relevant Terms

which were then used in clustering. Rahgozar and Inkpen (2016a) used supervised learning to classify Hafez. We tried an unsupervised method and did not use master-labels by Houman (1938) as training, but we used his labels to evaluate our clusters. For a long time, researchers tried to extract what was implied in the context, by applying generative models and collocation of the words. For example Brown et al. (1992) assumed word clustering carried semantic groupings. Our corpus was considerably smaller than those in the literature, none-the-less, hand-labeling or human annotation is an expensive, rare and slow process. Therefore, similar to many NLP researchers, we used clustering to augment annotated data based on the assumption that word clusters contained specific semantic information (Miller et al., 2004). Capturing semantic latent properties has been a long and continuous effort in Computational Linguistics. (Deerwester et al., 1990) used *singular-value decomposition* as pseudo-document vectors to detect implicit semantic properties, referred to as latent semantic analysis (LSA) in text. This was what we intended to do but in poetic text. In the continuation of semantic endeavour, (Blei et al., 2003) later developed latent Dirichlet allocation (LDA), an unsupervised generative probabilistic model to extract topics and their important associated terms. We used LDA driven features, before passing them as vectorized corpus to the K-Means clustering algorithm. Inkpen and Razavi (2013) used LDA driven features for semantic classifications of news group texts. Asgari et al. (2013) used topic models (unsupervised learning) to cluster Persian poetry by genre and then compared the results with SVM (supervised learning) classifications. Similarly, we used latent semantic indexing (LSI) and LDA-driven features for clustering. Saeedi et al. (2014) also used unsupervised semantic role labeling in Persian, but used different clustering scores than ours, such as purity and inverse-purity. We also used word embedding as features (Mikolov et al., 2011), which was the basis of our challenger model, against the top champion, the homothetic model. Zhang and Lapata (2014) used word embedding in poetry generation task and found it an effective feature for capturing the context.

The concept of *similarity*, mostly translated to *distance* in mathematics, is inherent and fundamental, especially in clustering and unsupervised learning algorithms. Kaplan and Blei (2007) for example, used vector space and principal components analysis (PCA), to depict style similarities in American poetry. Correlation was also used as a similarity measure to detect topics in poetry (Asgari and Chappelier, 2013). Lee et al. (2005) concluded that measures such as correlation, Jaccard and Cosine similarities performed almost the same in clustering documents. Similar to our research, Chambers and Jurafsky (2009) used but chain-similarities in an unsupervised learning algorithm, to determine narrative schemas and participants of semantic roles, instead of relying on any hand-built classes or knowledgebase. Their similarity definition was based on a pairwise summation of PMI and Log-Frequency of their narrative schema's vector representations. Then they maximized those similarities to score and determine semantic-role labels. Herbelot (2014) used similarity of word distributions, in pursuit of detecting semantic coherence in modern and contemporary poetry.

6 Conclusion

Capturing semantic attributes of text by machine learning has been an open research area. Houman's (1938) chronological and semantic classification of Hafez, unique up to now, assumed the young poet had a different world-view than the old, hence the difference would be reflected in his poetry, in terms of meaning. We created the first series of unsupervised semantic classifications of Hafez; using LDA, LSI, Log-

Entropy, Doc2Vec and similarity-driven features to capture such nuances of meaning. We showed that these NLP tools could help to produce different clusters of poems, to complement their scholarly hand-labeled version. We introduced the similarity-based features to build our champion models. We observed that our homothetic clustering had a slightly higher homogeneity, completeness and much better silhouette scores compared with our other features, but kappa distribution with Houman labels, was not statistically significant. Yet, in the analysis of our homothetic clustering results, we could trace the effect of similarity to the anchor poems. In case of HEP for example, clusters seemed to be more "aware" of classes "Youth" and "Senectitude", from which the anchors had been chosen.

Using LSI and LDA-driven features, similar to those Rahgozar and Inkpen (2016b) proved effective in chronological classification of Hafez poems, plus other semantically effective features, we created new sets of labels, not necessarily chronological, yet semantically different.

We applied our top homothetic feature engineering that proved the most effective in our clustering, to predict the whole Hafez corpus as a parallel labeling to Houman's. We investigated semantic differences, using both labels while comparing and tracing the consistencies through visualizations. We developed rigorous semantic analysis, refined and guided our homothetic clustering framework to get closer to Houman's ground-truth if possible. We provided multiple perspectives by our automatic labeling results and framework to support semantic analysis in literary scholarship.

6.1 Results

- Doc2Vec-P word-embedding scored higher coherence[20] and silhouette than other non-homothetic features used in Hafez automatic clustering experiments;

- We created two new sets of automatic labeling for Hafez corpus, by Doc2Vec as challenger and Sim^2 as champion clusterers, which had 0.58 *kappa* and 0.86 correlations but had insignificant resemblance with the Houman labels, 0.034 kappa at best(HRP-6cls);

- Sim^2 did not fully qualify as a quasi-semi-supervised[21] algorithm, given the low linear kappa with Houman, but proved to be a powerful clusterer, reaching (high coherence and) silhouette scores, of up to 95%;

- Sim^2 was the only clusterer to perform at its best with 6 clusters, equal to Houman classes, $k = cls$;

- None of the automatically generated labels were showing signigicant consistency with Houman's classification, but provided with new semantic perspectives to Hafez studies;

- Semantic evaluations and visulaizations helped validate the clustering results, using random poems;

- Visualizations in conjunction with homothetic clustering could be used to build a poetry analysis tool to support literary scholarship and research, even with small corpora such as ours.

Inspired by Houman's (1938) semantic approach, one can replicate and apply our poetry clustering framework to other poetic texts, as a means of assisting and enabling literary research and scholarly analysis of poetic text by clustering. We have also made the results of our clustering and new labels available for literary research and public use. Our guide is with refernce to the Houman's order of poems, which is based on Ghazvini copy[22] (see Appendix A).

References

Ehsaneddin Asgari and Jean-Cédric Chappelier. 2013. Linguistic resources and topic models for the analysis of persian poems. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 23–31.

Ehsaneddin Asgari, Marzyeh Ghassemi, and Mark Alan Finlayson. 2013. Confirming the themes and interpretive unity of ghazal poetry using topic models. In *Neural Information Processing Systems (NIPS) Workshop for Topic Models*.

[20] *Coherences* were not reported here specifically as they were reflected in *Silhouette* scores by definition.

[21] Handpicked anchors did not significantly increase kappa with Houman labels.

[22] An old reliable source of Hafez poems.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics.

Jason Chuang, Christopher D Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces*, pages 74–77. ACM.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Aurélie Herbelot. 2014. The semantics of poetry: A distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Mahmoud Houman. 1938. *Hafez*. Tahuri.

D. Inkpen and A. H. Razavi. 2013. *Topic Classification using Latent Dirichlet Allocation at Multiple Levels*. School of Electrical Engineering and Computer Sci. University of Ottawa.

David M Kaplan and David M Blei. 2007. A computational approach to style in american poetry. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 553–558. IEEE.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Michael D Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. 2011. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Arya Rahgozar and Diana Inkpen. 2016a. Bilingual chronological classification of hafez's poems. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 54–62.

Arya Rahgozar and Diana Inkpen. 2016b. Poetry chronological classification: Hafez. In *Canadian Conference on Artificial Intelligence*, pages 131–136. Springer.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Parisa Saeedi, Heshaam Faili, and Azadeh Shakery. 2014. Semantic role induction in persian: An unsupervised approach by using probabilistic models. *Literary and Linguistic Computing*, 31(1):181–203.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Carl P Simon and Lawrence Blume. 1994. *Mathematics for economists*, volume 7. Norton New York.

Kai Zhang, Liang Lan, James T Kwok, Slobodan Vucetic, and Bahram Parvin. 2015. Scaling up graph-based semisupervised learning via prototype vector machines. *IEEE transactions on neural networks and learning systems*, 26(3):444–457.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680.

A Appendix

The most reliable print of Hafez is by Ghazvini, in which poems are organized alphabetically. The mapping table of the alphabetical order of poems to Houman classification can be found in (Houman, 1938).