Bilingual Chronological Classification of Hafez's Poems

Arya Rahgozar & Diana Inkpen
SEECS, University of Ottawa
Ottawa, Ontario, Canada
{arahg096,Diana.Inkpen}@uottawa.ca

Abstract

We present a novel task: the chronological classification of Hafez's poems (ghazals). We compiled a bilingual corpus in digital form, with consistent idiosyncratic properties. We have used Hooman's labeled ghazals in order to train automatic classifiers to classify the remaining ghazals. Our classification framework uses a Support Vector Machine (SVM) classifier with similarity features based on Latent Dirichlet Allocation (LDA). In our analysis of the results we use the LDA topics' main terms that are passed on to a Principal Component Analysis (PCA) module.

1 Introduction

Chronological classification of any artwork is a worthwhile task. We focus on the poetry of the giant of Persian poetry, Hafez from Shiraz. The purpose of our automatic chronological classification of Hafez's ghazals is to establish the relative timing of any poem concerning Hafez's lifetime, and thus to help understand his poetry better, while applying a semantic analysis approach. The objective of this research is to classify ghazals using machine learning (ML) techniques with scholarly benefits rooted in literary analysis and hermeneutics.

Harsh political conditions of Hafez's time required a unique type of encryption and mystical quality to the poems. As a result, scholars have argued for centuries about the ghazals' possible interpretations and engaged in enduring polemics over the subject.

We draw on the work of an outstanding author, Dr. Mahmood Hooman. In his seminal book about Hafez from about 80 years ago (Hooman, 1938), he has partially done this chronological classification by hand.

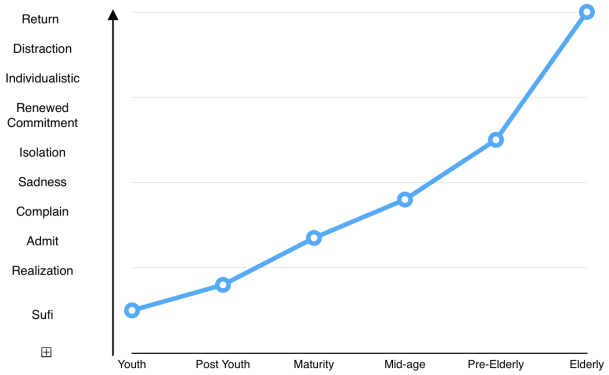

Figure 1: Hafez's Evolutionary Growth Curve

Hooman provides a psychological and personality-growth perspective on the poet Hafez. This perspective plays an integral role in the interpretation of the poems and their chronological classification – see Figure 1. This analytical spectrum of Hafez and his ghazals has been our guidance in deciding to apply Natural Language Processing (NLP) semantic-based methods in the chronological classification of Hafez's ghazals.

We considered the task as a deserving candidate for automatic text classification by ML. From the very beginning, we realized the great challenges involved. Most important, there was no large and reliable corpus of Hafez poems available in electronic form. Therefore we built one composed of all the 468 ghazals, each about 10 lines. We were able to include good English version only for 71 of them.[1]

In addition to classification, we also decided that we need some means of providing an intuitive rationale for each prediction. Therefore, in the end, we applied a Topic-Term analysis to the poems to address that.

[1]English translations are by Shahriar Shahriari.

Proceedings of the Fifth Workshop on Computational Linguistics for Literature, NAACL-HLT 2016, pages 54–62,
San Diego, California, June 1, 2016.

2 Hafez Corpus

We have used Ghazvini's version of Hafez's poems[2] and we followed Hooman's approach. We have also added the English translations whenever available. While typing up the poems, we applied predefined rules to all poems. In other words, we ensured consistency while creating our Hafez corpus. It is one of the attributes of an ancient language such as Persian to be flexible, and to provide freedom and variety of writing options within the same compound terms. This variety comes at the expense of complex computational implications. We had to apply consistency rules so that any current or future parsing of the terms is consistent across all 468 ghazals. We have used multiple types of white spaces to separate or join the one-word terms that we write as counter-intuitive in Persian.[3] In places of potential confusion, we have specified the otherwise unwritten vowels and diacritics inline.

As we see in Figure 1 in the chronological and conceptual poem chart, each poem essentially would reside on a specific curve point depending on its determined point in time, and on its semantic elements, theme and attributes that Hooman detected in the poems. Our corpus follows exactly Hooman's order of ghazals.

We have derived rules from the Persian linguistics, defined procedures and specifications, and applied them to our Persian corpus during its development. From the 468 ghazals, Hooman labeled only 249 with time information. We have consolidated six classes of chronological pairs into three (Youth, Maturity and Senectitude) to facilitate classification experiments, as shown in Table 1 (combining labels a and b into a', c and d into b', and e and f into c').

Table 1: Corpus Training Labels

Six Classes		Three Classes
Youth = 38	a	a'
After Youth = 25	b	
Maturity = 79	c	b'
Middle Age = 66	d	
Before Senectitude = 28	e	c'
Senectitude = 13	f	

3 Related Work

The Cross-Language Text Categorization (CLTC) task often concerns categorizing text based on the labeled training data from one language to help to classify text in another language. Popular techniques use the bag-of-word (BOW) method as a base to classify texts. Researchers obtained varied high accuracies in text classification depending on the task, context and corpus size. One source of differences is in how features are developed and weighted; another is in the learning algorithms. Gliozzo and Strapparava (2006) built common etymological ancestry attributes of words between Italian and English, which were used to train an SVM model in one language to classify text in the other. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) was used to create a deep vector representation of the word-document co-occurrences of shared lexical and etymological attributes. Dumais et al. (1997) found semantic correspondences between languages by using LSA and SVM to create multilingual domain models.

Languages adopt words from each other and adjust them for their purposes, yet maintain their common roots. For example, the words *Check*, *Chess* and *Checkmate* in English correspond to their Persian roots as *Shah* and *mat* in *Shah-mat*. In this way, the two languages preserve strong semantic relations.

Luštrek (2006) has a good discussion and overview of the types of features used in text classification, with a focus on genre detection in text classification. Simonton (1990) present experiments in authorship attribution for poetry analysis and lyrics using shallow features such as part-of-speech (POS) features and function word distribution. Simonton analyzed the 154 sonnets attributed to William Shakespeare. Each sonnet was partitioned into four consecutive units (three quatrains and a couplet), and then a computer tracked down how the number of words, different words, unique words, primary process imagery and secondary process imagery changed within each sonnet unit. He no-

[2]Mohammad Ghazvini (1874-1949), an Iranian scholar, corrected and prepared today's most reliable prints of Hafez ghazals.

[3]For example, dâneʃ-âmuz 'student' is one word, but we write it as two in Persian.

ticed a common vocabulary change in the end unit, the couplet. Kim et al. (2011) used deeper features such as the distribution of syntactic constructs in prose to analyze authorship and writing style. Synonyms and hyponyms are also used as features (Scott and Matwin, 1998). The POS proportion of *hapax legomena* per document plus end of line rhyme have been examined as features (Mayer et al., 2008). Hirjee and Brown (2009) showed that a statistical rhyme detector can extract in-line slant rhymes to analyze Rap lyrics.

To approximate publication time of the lyrics and detection of the genre, Fell (2014) used features such as vocabulary, style, semantics, orientation and structure of the song for an SVM classifier.

According to (Zrigui et al., 2012), an LDA-SVM model is the best performing classifier in finding main subject heading of Arabic texts; they compared this top performer with Naive Bayes, SVM and kNN classifiers. Luo and Li (2014) employ a two-phased LDA-SVM model to classify about 20 different newsgroup texts. They used LDA, Probabilistic Latent Semantic Indexing (PLSI), PCA, Hierarchical LDA and SVM to classify such documents.

Razavi and Inkpen (2014) used SVM with multilevel LDA features to classify social media messages and newsgroup texts. In search of an efficient text classification method and following the related works mentioned above, we decided to use SVM (Cortes and Vapnik, 1995), because it is a state-of-the-art classification algorithm (Joachims, 1998).

Orthographic, syntactic and phonemic features were used to classify poems by style (Kaplan and Blei, 2007). In analyzing poems and their aesthetics to reach the semantics of imagery, other researchers employ sound devices such as alliteration, consonance and rhyme (Kao and Jurafsky, 2015). More work uses NER and POS taggers to create features to classify poems by style (Delmonte, 2015). Lou et al. (2015) classified poems into nine classes (Love, Nature, Religion and other), allowing a poem to be in more than one class.

Unlike the previous work on poetry classification, we classify the poems by one poet alone – Hafez – in chronological order, and the poems contain many symbols and hidden semantics that we captured by LDA-driven cosine similarities in vector space.

4 Proposed Methodology

As shown in Figure 2, we used feature-engineering techniques based on Bag-Of-Words[4] and Term Frequency-Inverse Document Frequency (TF-IDF) that we transformed into the vector space of LSI or LDA. We then used those representations for training the SVM classifier. To get our best performing SVM classifier, we used a new representation based on cosine similarity measures calculated from LDA topics. The dictionary maps a poem's normalized words into an index.

Similarly to the work of Bezdek et al. (1998) inspired by (Chang, 1974), we have averaged *Prototype* similarity vectors for each class. That is, each poem has three *Prototype* features to train SVM. We first calculated each poem's similarity to others then we averaged that by class. In other words, each of the three features is the ghazal's average LDA-driven *cosine similarity* to all other poems of each class, calculated one by one, to capture their probabilistic semantic relatedness.

We discuss in section 6 the highest probability terms among all six[5] topics for each class – *Youth*, *Maturity* and *Senectitude*– to analyze the results. We used the GENSIM library (Řehůřek and Sojka, 2010) to develop the features; the similarity features in GENSIM and its indexing mechanism by LSI concepts are based on (Deerwester et al., 1990). Then we use WEKA (Hall et al., 2009) to train the SVM classifier. We grouped the six classes of Hooman into three, for performance reasons. In the source data, wherever available, English translations are directly appended to the poems' Persian instances. Similar to Figure 3, that shows the LDA clusters for each class (only one term from each cluster is shown), we also created the cluster of top terms for predicted poems for error analysis purposes. We compared the associated class terms with those for each predicted class of ghazals to study the internal topic attributes and hence we were able to provide clues for predictions. We hope that the results of our analysis will help NLP researchers to both observe the effects of LDA topic terms in liter-

[4]The frequency of each word used as a feature, irrespective of grammar, order or semantic relations.

[5]More LDA topics did not produce any important lift in performance.

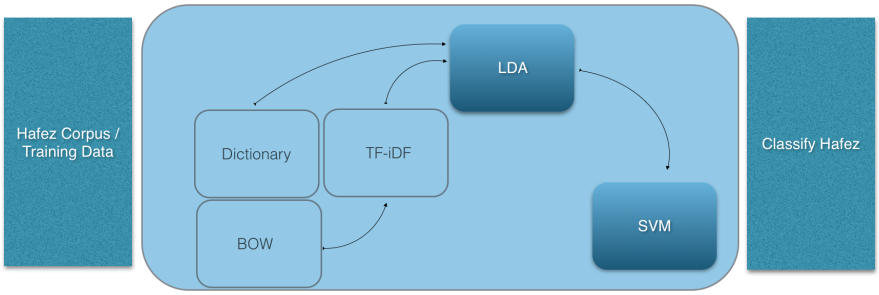

Figure 2: Technical High-level Process

ature contexts and to extend our insight further over the poems of Hafez.

As part of the final analysis of the results, we have used PCA to reduce dimensionality and to draw the LDA results in 2D for analysis purposes. Driven by the LDA model, clusters of words may slightly differ in each run. We were able to show that LDA terms relations by PCA, bring about consistency, relatively maintain comparability of distinctive characteristic of a ghazal and its class for which the prediction is made, and therefore help the user not only better distinguish between the ghazals possible classes but also better justify the classification theme on Hooman's classes. Kaplan and Blei (2007) use PCA so that they can visualize similarity among poems. They use orthographic, syntactic and phonemic features to tackle, distinguish and classify poems by style. Kao and Jurafsky (2015) extended that work and introduced other features of sound devices – such as alliteration, consonance and rhyme – in order to analyze further poems and their aesthetics to get at the semantics of imagism. The cluster of terms caused interesting discussions amongst our experts. Term clusters also played a critical role in providing the rationale for the predictions and their comparisons and interpretations.

5 Experiments and Results

The baseline accuracy for the classification of the three amalgamated classes is 58.2.%[6]

In Table 2, we show the results of tenfold cross-validation for our SVM classifiers with different sets of features. The evaluation measure is the weighted average of the F-measures proportional to the number of elements in each of the three classes, as calculated by WEKA.[7]

Table 2: SVM Classification Results for 3 classes (F-measure)

Features	Language	
	Persian	Persian-English
BOW	61%	65.1%
LDA	56.2%	58.2%
BOW+LSI	61.4%	65.1%
BOW+LDA	61.8%	65.1%
LDA Similarity	79.52%	78.4%

In our first experiment, we created the BOW training data as input to the SVM classifier and increased the F-measure to 61%. The LDA factors alone did not go above the baseline of 58%. Keeping the BOW and adding the LSI or LDA factors only slightly improved the F-measure over the BOW alone. A t-test showed a 95% confidence that the results improved significantly when we added the English translations.

At that point, we hypothesized that the LSI- or -LDA-driven similarity factors alone should provide us with strong enough training features. Therefore, in the next experiments, we went back and created the SVM training data only with normalized similarity factors, once with LSI and once with LDA. LDA driven similarity factors proved stronger than those of LSI. That is, as we observed the remarkable strength of these features, we only kept the BOW and LDA factors in the similarity factor calculations, in the final SVM training data. Yet this method brought the accuracy of the classifier to our best result of 79.5% using our Persian training dataset.

[6]The Baseline is a classifier that always chooses the most frequent class, b', out of the three.

[7]`http://weka.sourceforge.net/doc.dev/weka/classifiers/Evaluation.html`

The English addition, only in this case, reached a plateau. That, we believe, was due to the scarcity of the features – only three.

To analyze the errors made by the classifier, we looked at the confusion matrix with columns showing the "classified as". We noticed that the classifications faults were often caused by classes a' and c', which make up the smaller sections of the corpus; they are under-represented:

a'	b'	c'	
44	19	0	a'
0	145	0	b'
17	15	9	c'

We have also used the trained model for predicting the classes of the unlabelled ghazals. We then asked our two experts, who consistently validated the labelling results, for a few of the unlabelled ghazals.

6 Analysis of the Results

We only discuss the main term from each of the 6 LDA topics of each class.[8] In the next analysis, we will look at how they correspond each of the top LDA terms of a sample poem from each class. For brevity, we only show one poem per class but in fact the framework proved useful in providing us with insightful clues and consistent intuitive reasoning behind the classifier predictions.

6.1 First Period - Youth

The Youth class has the following cluster of terms:[9]

0. Vision *nazar*, Connected *vasl*, Unable *nâtavan*, Complain *ʃekâyat*, Your Sorrow *qamat*, A Heart *deli*, Glass *ʃɪʃə*, Repentance *tobæ*, Universe *jahan*, Hand *dast*.
1. Other *degar*, Flower *gol*, Remenisance *bovad yâd*, Airy *havâi*, Solution *tadbɪr*, Jam *jam*, Wine *məɪ*, Guru *pɪr*, Hand *dast*.
2. Is *ast*, From *ke az*, Sorrow *qam*, Be *bâʃ*, There *ânja*, In *andar*, Blood *xʊn*, Wine *mɛy*, Full *por*, To Be Me *bâʃam*.
3. Arch *tâq*, Gem *laæl*, Because *bahrɛ*, You *to*, Face *didɛ*, Speech *ʃirin-soxan*, Limit *hadd*, Business *kâr*, No Hint *nemibinam-neʃân*, Ruined *xarâb*.
4. Secret *sərr*, Destiny *qadar*, Say *gʊ*, Cup *jâm*, Know *dânɪ*, Friends *ʏârân*, Came *âmad*, Dawn *sahar*, Life jân.
5. Break *beʃkan-bɛ*, Title *maqâm*, Life *jân*, Thousands *ʃiezârân*, Loose *sost*, Candle *ʃamæ*, My Heart *delam*, Love *əʃq*, Downhill *naʃib*.

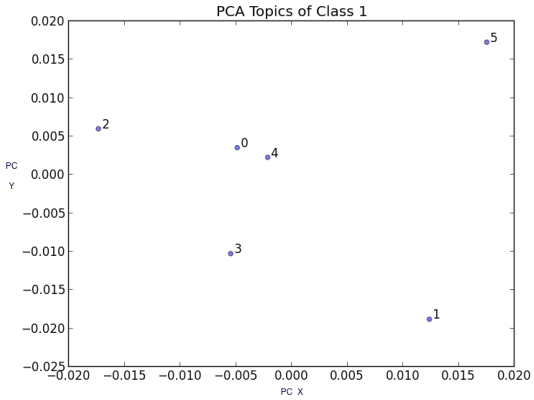

Figure 3: LDA topics for the class Youth

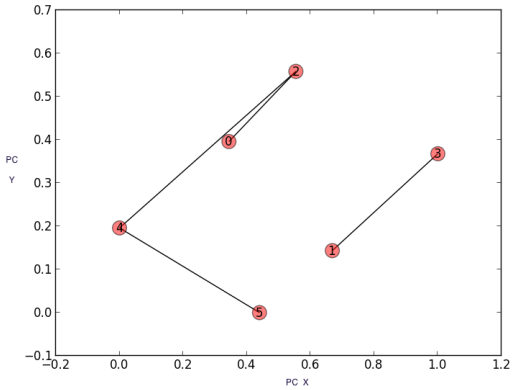

Figure 4: LDA Topics; Graph Relations for the class Youth

6.1.1 Analysis of poems: Class Youth

Let us look at a poem that Hooman has classified as belonging to the Youth period of Hafez's life (and for which the prediction was also the Youth class). Let us observe what elements and cluster of words we see in the ghazal. Here is the first line of the ghazal 48:[10]

[8] We experimented with multiple LDA topic numbers but here we only show the results for the Six-Topic top terms for each class and each individual ghazal for comparison.

[9] Persian words in the phonetic form are in italics.

[10] It is 48 according to Hooman's numbering system.

sahargah rahrovi dar sarzamini - hami goft in moamma ba qarini
and the translation of the ghazal is as follows:
A traveler in a strange land Took a stranger by the hand
You will only see clarity of the wine If for forty days you let it stand.
God keep us from the dervish's cloak That conceals an idol in every strand.
Though virtue needs no recognition Let helping the needy be your errand.
O you the owner of the harvest Keep your harvesters from reprimand.
Where has all the joy gone? Why is the pain of love so bland?
Every chest is gloomy dark and sad; Let love's flame in hearts be fanned.
Without the finger of lovers For golden rings there's no demand.
Though Beloved seems to be so harsh The lover accepts every command.
Walk to the tavern and I will ask Have you seen the end you have planned?
Neither Hafiz's heart is in lessons so grand Nor the teacher can fully understand.

By looking at this ghazal,[11] one can observe that the terms *Glass*, *Heart* and *Sorrow* correspond with topic 0. *Sorrow* also belongs to topic 2, but from topic 2, we also have *Is* occurring twice and *Be* is present 5 times. Interestingly enough, the network in Figure 4 shows that there is a relationship between topics 0 and 2. Elements of topics 1 and 5 are depicted as far from topics 2 and 0 in the PCA chart, and accordingly they are not present. Overall, the elements and genre of the ghazal are very consistent with the concepts depicted by the word clusters and topic charts of this class.

As we observe in Figure 3, topics 1, 2 and 5 are the farthest from each other, but in the network or the weighted-Euclidian-distance Figure 4, topics 1 and 3 have no relations with others in the graph. Topics 0, 2, 4 and 5 are related, in that order. These links indicate how the term characteristics of the topics interrelate. In this case, we are more likely to see topics 1 and 3 show up in a ghazal; but the cluster of words in topics 1 and 2 are hardly expected to show up in the same ghazal. One can also observe the contrast between the two topics 1 and 2; that is, the topic 1 is obviously more positive than topic 2.

[11] We generated the topic terms using the Persian corpus, so the exact term may not necessarily exist in this poetic translation by Shahriar Shahriari.

6.2 Second Period: Maturity

The Maturity class has the following cluster of terms:

0. Objective *hâjat*, Dust *xâk*, Hafez *hâfez*, Grace *mənnat*, Excited *barafruxtəh*, Palate *kâm*, Heart *del*, That *kə*, Concern *kâr*.
1. Vision *nazar*, Life *jân*, Return *bɑz*, Universe *jahan*, Cleanliness *taharat*, Is *st*, Secret *serr*, So *ke*, Is *ast*.
2. Hafez *hâfez*, Heart *del*, Soleiman *soleimân*, Virtue *honar*, Word *soxan*, Distressed *parijân*, See *bin*, Where *kojâ*, Candle *ʃamæ*, Vision *nazar*.
3. Went *raft*, Return *bâz*, Not Remain *namânad*, Flower *gol*, You *to*, Sweetheart *yâr*, When *kəy*, Harm *balâ*, Sympathy *deli*.
4. Envy *hasrat*, and *va*, Said *goftâ*, That *kæ*, Dust *xâk*, This way *kɛ-ɪn*, Cup *jâm*, Palate *kâm*, come I said *âyad-goftam*, Come *biâ*.
5. I want *xâham*, Has Left *nahâdæ*, Cannot *natavân*, Wrong *qalat*, Eye *ɕaʃm*, Contract *ahd*, Is-Not *nist*, Wine *məy*.

6.2.1 Analysis of poems: Class Maturity

An example of analysis of this section is ghazal 206 of Hooman's classification labeled Maturity. The first line of this ghazal starts with this: *salha dafter ma dar geroye sahba bʊd - ronaghe meikade az darso daaye ma bʊd.*

The translation of the ghazal is as follows:
For years to the red wine my heart was bound The Tavern became alive with my prayer and my sound.
See the Old Magi's goodness with us the drunks Saw whatever we did in everyone beauty had found.
Wash away all our knowledge with red wine Firmaments themselves the knowing minds hound.
Seek that from idols O knowing heart Said the one whose insights his knowledge crowned.
My heart like a compass goes round and round I'm lost in that circle with foot firmly on the ground.
Minstrel did what he did from pain of Love Lashes of wise-of-the-world in their bloody tears have

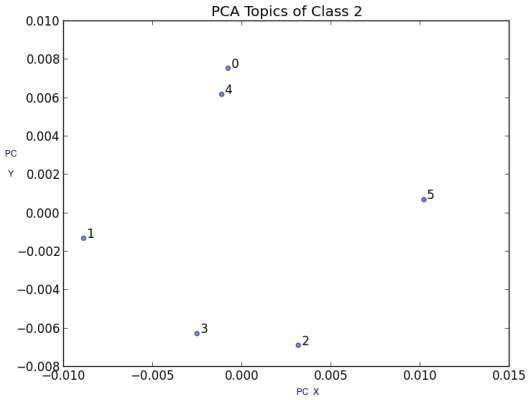

Figure 5: LDA topics for the class Maturity

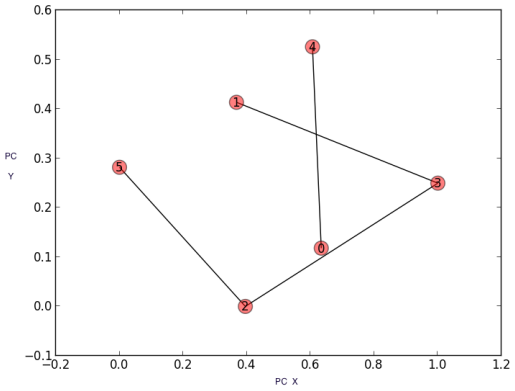

Figure 6: LDA Topics, Graph Relations for the class Maturity

drowned.
With joy my heart bloomed like that flower by the stream Under the shade of that tall spruce myself I found.
My colourful wise Master in my dealings with the black robes My meanness checked and bound else my stories would astound.
Hafiz's cloudy heart in this trade was not spent This merchant saw and heard every hidden sight and sound.

We observe in Figure 5 that the highest number of terms in the term cluster belongs to topics 0 and 4. The term *That* occurs 5 times and 2 times in slightly different form, *Heart* and *Said* are common on topics 0 and 4. The system depicted this relation by both the topics chart and network distance relation charts. Then we observe the terms *Vision* and *Universe* from topic 1 and the term *See* from topic 2 and the term *Flower* from topic 3, each once. Network Figure 6 depicts this relation.

6.3 Third Period: Senectitude

The Senectitude class has the following cluster of terms:

0. Prescription *davâ*, Universe *donyâ*, Does it *bekonad-ze*, Wonder *ajab*, Happy *xoʃ*, Kindness *məhr*, Cup *jâm*, Is *bovad*, Veil *hejâb*, Free *rahâ*.
1. Life *jân*, Song *âvâz*, That from *kaz*, Scream *faryâd*, All *ɦamæ*, In *andar*, Nightingale *bolbol*, Universe *jahân*, Let it become *ʃavad*.
2. Full *por*, And *va*, That-this *kɪn*, Sadness *qam*, That *ke*, Became *bəʃod*, Witness *ʃâhed*, Wine *mɛy*.
3. Word *soxan*, Sun *xorʃid*, Can *tavâni*, Is Not *nabovad*, Light *ɕerâq*, Is going *miravad*, Monastry *somæ*, Nice *nekʊ*, Is not *st-na*, You *to*.
4. Fell off *oftâd-az*, Fell *oftâd*, My heart *delam*, Blood *xʊn*, Does from *konad-zə*, Hand *dast*, Universe *jahân*, Love *əʃq*, Familiar *ahl*, Smell *bʊyə*.
5. Better *behtar*, Wisdom *aql*, Turn *nɔbat*, Is *st*, Drink *bâdeh*, Within *andar*, To *râ*, Wine *məy*, From that *kaz*.

6.3.1 Analysis of poems: Class Senectitude

We have randomly chosen ghazal 241, which Hooman classifies into the last class. It starts with the line: *har chand piro khaste delo natavan shodam - har gah ke yɑde rʊye to kardam, javɑn shodam.*

The translation of the poem is as follows:

Though I am old and decrepit and weak My youth returns to me every time your name I speak.
Thank God that whatever my heart ever desired God gave me that and more than I ever could seek.
O young flower benefit from this bounty In this garden I sing through a canary's beak.
In my ignorance I roamed the world at first In thy longing I have become wise and meek.
Fate directs my path to the tavern in life Though many times I stepped from peak to peak.
I was blessed and inspired on the day That at the abode of the Magi spent a week.
In the bounty of the world await not your fate I found

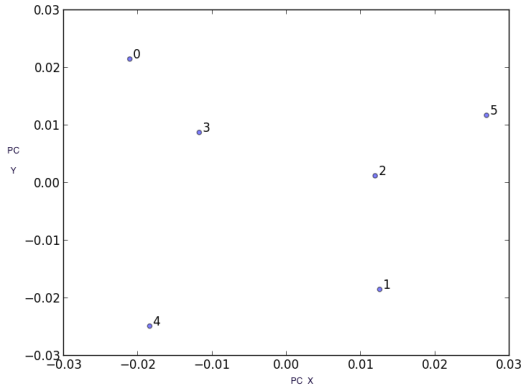

Figure 7: LDA topics for the class Senectitude

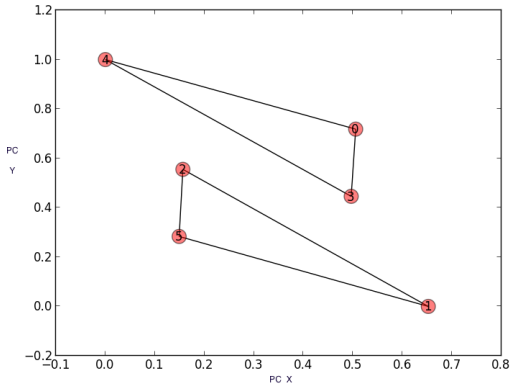

Figure 8: LDA Topic, Graph Relations for the class Senectitude

the Beloved when of wine began to reek.
From the time I was entrapped by thy eyes I was saved from all traps and paths oblique.
The old me befriended the unreliable moon Passage of time is what makes me aged and weak.
Last night came good news that said O Hafiz I forgive all your errs even though may be bleak.

Obviously, this is in agreement with Hooman's descriptions of the attributes of this class, as it shows a very introvert and sad poet who has fewer connections with this natural world. It has specific mentions of Hafez referring to himself as being old. Let us see how our developed cluster of terms plays out in this case.

Consistent with Figure 7, although we see the sporadic presence of nearly all topic terms except that of topic 5, we see that topic 2 is dominant, as we observe and identify the associated cluster of terms of this group such as *That*, three times, *Sad* and *Wine*. The next topic is topic 1 with the terms *Nightingale*, *Universe* and *That-From*.

We observe the terms *My Heart* and *Universe* of topic 4, *Cup* of topic 0 and *Wine* and *Is* of topic 5; but *Wine* also overlaps with topic 2. The term *You* of topic 3 shows up three times.

The interesting symmetric nature of the relation network Figure 8 for this class is consistent with our observation that we have a strong presence of the clusters of terms 1, 2 and 5 and a weaker presence of topics 0, 3 and 4 in which the term *Universe* is common! If we exclude the term *Universe* from both sub-graph terms, looking at the network Figure 8, there are only three distinct terms in the weaker group (0,3,4) compared to the other sub-graph (1,2,5) with term presence of 7.

7 Conclusion and Future Work

Our model automated the classification of Hafez's ghazals using our Hafez corpus. We used LDA and SVM to detect the semantics of Hafez's ghazals and to classify them chronologically. In future work, we will use automatic translations, and will add features such as word embeddings to improve the classification. We are planning to add word-embedding features to enhance our training data and try it with all six Hooman classes. We hope that by increasing the number of features in the training data set we can further improve the more granular classification performance. PCA helped with the intuitive analysis and validation of the prediction results but it deserves a whole paper dedicated to it. We will present more rigorous visualization of topic term validation methods. Another direction of future work is to automatically detect earlier poets' style and rhythms, given the fact that Hafez represents the apex of Persian poetry after *sa:di*, xâqâni, *dehlavi* and others. Ashoori (2011) strongly believes that we can even find obvious influences of important books such as *mersad-ol-ebad* and *kaʃfol-asrâr*. It would be worthwhile to use ML to draw relations, detect and rank such traces in Hafez's poetry and its hermeneutics.

Acknowledgments

Heartfelt thanks to Mr. Mehran Rahgozar for his continuous expert advice, comprehensive support

and considerations with the preparation of the Hafez corpus and implementation of its linguistic properties and literary evaluation of predicted poems. Our special gratitude also extends to Mr. Mehran Raad for his inspiring literary conversations about Hafez and expert advice in the evaluation of the results. We thank the reviewers for their most helpful comments.

References

Dariush Ashoori. 2011. *Erfan o Rendi in Hafez Poetry*. BBC Interview.

James C Bezdek, Thomas R Reichherzer, Gek Sok Lim, and Yianni Attikiouzel. 1998. Multiple-Prototype Classifier Design. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 28(1):67–79.

Chin-Liang Chang. 1974. Finding Prototypes for Nearest Neighbor Classifiers. *Computers, IEEE Transactions on*, 100(11):1179–1184.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine learning*, 20(3):273–297.

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.

Rodolfo Delmonte. 2015. Visualizing Poetry with SPARSAR–Visual Maps from Poetic Content. *Workshop on Computational Linguistics for Literature*, pages 68–78.

Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic Cross-language Retrieval Using Latent Semantic Indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, pages 15–21.

Michael Fell. 2014. Lyrics Classification. Master's thesis, Saarland University.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting Comparable Corpora and Bilingual Dictionaries for Cross-language Text Categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 553–560. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA Data Mining Software: an Update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hussein Hirjee and Daniel G Brown. 2009. Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics. In *ISMIR*, pages 711–716.

Mahmood Hooman. 1938. *Hafez*. Tahuri.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Claire Nédellec and Céline Rouveirol, editors, *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin. Springer.

Justine T Kao and Dan Jurafsky. 2015. A Computational Analysis of Poetic Style. *LiLT (Linguistic Issues in Language Technology)*, 12(3):1–33.

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.

Sangkyum Kim, Hyungsul Kim, Tim Weninger, Jiawei Han, and Hyun Duk Kim. 2011. Authorship Classification: a Discriminative Syntactic Tree Mining Approach. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 455–464. ACM.

Andres Lou, Diana Inkpen, and Chris Tanasescu. 2015. Multilabel Subject-Based Classification of Poetry. In *The Twenty-Eighth International Flairs Conference*.

Le Luo and Li Li. 2014. Defining and Evaluating Classification Algorithm for High-dimensional Data Based on Latent Topics. *PloS one*, 9(1):e82119.

Mitja Luštrek. 2006. Overview of Automatic Genre Identification. *Ljubljana, Slovenia: Jožef Stefan Institute, Department of Intelligent Systems*.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. 2008. Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *ISMIR*, pages 337–342.

Amir H Razavi and Diana Inkpen. 2014. Text Representation Using Multi-level Latent Dirichlet Allocation. In *Advances in Artificial Intelligence*, pages 215–226. Springer.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *LREC2010 Proceedings*. University of Malta, May.

Sam Scott and Stan Matwin. 1998. Text classification using WordNet Hypernyms. In *Use of WordNet in natural language processing systems: Proceedings of the conference*, pages 38–44.

Dean Keith Simonton. 1990. Lexical Choices and Aesthetic Success: A Computer Content Analysis of 154 Shakespeare Sonnets. *Computers and the Humanities*, 24(4):251–264.

Mounir Zrigui, Rami Ayadi, Mourad Mars, and Mohsen Maraoui. 2012. Arabic Text Classification Framework Based on Latent Dirichlet Allocation. *CIT. Journal of Computing and Information Technology*, 20(2):125–140.