Automatic Poetry Classification and Chronological Semantic Analysis

Arya Rahgozar

Thesis submitted to the University of Ottawa
In partial Fulfillment of the requirements for the
PhD degree in E-Business
****

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

The correction, authentication, validation and identification of the original texts in Hafez's poetry among 16 or so old versions of his Divan has been a challenge for scholars. The semantic analysis of poetry with modern Digital Humanities techniques is also challenging. Analyzing latent semantics is more challenging in poetry than in prose for evident reasons, such as conciseness, imaginary and metaphorical constructions. Hafez's poetry is, on the one hand, cryptic and complex because of his era's restricting social properties and censorship impediments, and on the other hand, sophisticated because of his encapsulation of high-calibre world-views, mystical and philosophical attributes, artistically knitted within majestic decorations.

Our research is strongly influenced by and is a continuation of, Mahmoud Houman's instrumental and essential chronological classification of ghazals by Hafez. Houman's chronological classification method (Houman, 1938)[1], which we have adopted here, provides guidance to choose the correct version of Hafez's poem among multiple manuscripts. Houman's semantic analysis of Hafez's poetry is unique in that the central concept of his classification is based on intelligent scrutiny of meanings, careful observation the evolutionary psychology of Hafez through his remarkable body of work. Houman's analysis has provided the annotated data for the classification algorithms we will develop to classify the poems. We pursue to understand Hafez through the Houman's perspective. In addition, we asked a contemporary expert to annotate Hafez ghazals (Raad, 2019)[2].

The rationale behind our research is also to satisfy the need for more efficient means of scholarly research, and to bring literature and computer science together as much as possible. Our research will support semantic analysis, and help with the design and development of tools for poetry research.

We have developed a digital corpus of Hafez's ghazals and applied proper word forms and punctuation. We digitized and extended chronological criteria to guide the correction and validation of Hafez's poetry. To our knowledge, no automatic chronological classification has been conducted for Hafez poetry.

[1]Prof. M. Houman's book is available only in Persian. It is probably not available in the West. Please see Section 7.4

[2]Mr. Mehran Raad's labels were collected specifically for this thesis, and they are not available as a separate publication. Please refer to Section 7.4.

Other than the meticulous preparation of our bilingual[3] Hafez corpus for computational use, the innovative aspect of our classification research is two-fold. The first objective of our work is to develop semantic features to better train automatic classifiers for annotated poems and to apply the classifiers to unannotated poems, which is to classify the rest of the poems by applying machine learning (ML) methodology. The second task is to extract semantic information and properties to help design a visualization scheme to assist with providing a link between the prediction's rationale and Houman's perception of Hafez's chronological properties of Hafez's poetry.

We identified and used effective Natural Language Processing (NLP) techniques such as classification, word-embedding features, and visualization to facilitate and automate semantic analysis of Hafez's poetry. We defined and applied rigorous and repeatable procedures that can potentially be applied to other kinds of poetry. We showed that the chronological segments identified automatically were coherent. We presented and compared two independent chronological labellings of Hafez's ghazals in digital form, produced their ontologies and explained the inter-annotator-agreement and distributional semantic properties using relevant NLP techniques to help guide future corrections, authentication, and interpretation of Hafez's poetry. Chronological labelling of the whole corpus not only helps better understand Hafez's poetry, but it is a rigorous guide to better recognition of the correct versions of Hafez's poems among multiple manuscripts. Such a small volume of complex poetic text required careful selection when choosing and developing appropriate ML techniques for the task. Through many classification and clustering experiments, we have achieved state-of-the-art prediction of chronological poems, trained and evaluated against our hand-made Hafez corpus. Our selected classification algorithm was a Support Vector Machine (SVM), trained with Latent Dirichlet Allocation (LDA)-based similarity features. We used clustering to produce an alternative perspective to classification.

For our visualization methodology, we used the LDA features but also passed the results to a Principal Component Analysis (PCA) module to reduce the number of dimensions to two, thereby enabling graphical presentations. We believe that applying this method to poetry classifications, and showing the topic relations between poems in the same classes, will help us better understand the interrelated topics within the poems. Many of our methods can potentially be used in similar cases in which the intention is to semantically classify poetry.

[3]Our corpus consists of Persian and English so far.

Acknowledgements

I would like to offer my wholehearted and deepest gratitude and appreciation to Professor Diana Inkpen for her deep NLP knowledge, guidance, supervision, and for her confidence in me. It has been an invaluable and precious journey. Many thanks to Professor Inkpen for leading The Text Analysis and Machine Learning (TAMALE) group professionally, and managing seminars at which I got to learn plenty. Thanks to her teaching and support I got to know many scholars and researchers and their work. This thesis has been possible only because of her continuous guidance and support.

Many thanks to my Ph.D. defence panellists, Professor Bijan Raahemi, Professor Chris Tanasescu, Professor Ash Asudeh and Professor Stan Szpakowicz for their time, and for productive and professional feedback. Their invaluable critique and guidance have contributed to shaping this thesis in many ways.

Thanks to Professor Ash Asudeh for his mindful and productive remarks on the linguistic aspects of this work.

Special thanks to Professor Stan Szpakowicz for his unrivalled meticulous attention to detail and tireless care for the ultimate quality of research. It feels as if his force has been indirectly watching over me, making sure I get to the finish line successfully. I have been very privileged to reap the benefit of his presence all along!

I cannot thank everyone involved enough. I will never forget Professor Bijan's comments and productive remarks during my Ph.D. proposal defence session. For example, he mentioned that Hafez would provide me with sufficient material and that I could safely exclude Ferdowsi from this thesis. He would have known then that he would have probably saved my Ph.D. from never finishing. Professor Bijan's knowledgeable comments about machine learning along with his leadership helped me refine my research experiments and be organized.

Many thanks to Professor Chris Tanasescu (Margento) for his presence and interactions that have left me with a lasting feeling of kindness and confidence, along with a deep sense of appreciation for digital humanities. His remarks during the proposal and along the way have shaped the validity and directions of my graduate work.

I would also like to extend my preeminent sense of appreciation and respect to Professor Liam Peyton for his invaluable mentorship and support. He believed in me in so

many ways and provided me with so many educational and scholarly opportunities. He taught me how to be a professional researcher and scholarly leader and at the same be a decent, pure and caring educator.

My thanks to all members of the TAMALE group for their amazing interactions, openness and sharing throughout the past 6 years.

Many thanks to my father Mr. Mehran Rahgozar for his expert support in the preparation of the Hafez corpus and continuous linguistics discussions that inspired this work.

Thanks to Mr. Mehran Raad for his Hafez labels, expert support and inspiring scholarly discussions.

There is the only way in which I might be able to ever justify a bit why I have been subject to so much scholarly attention, professional support and pedagogical care I have received throughout my graduate studies at the University of Ottawa's Faculty of Engineering. I was surrounded by so many loving and pure-hearted, knowledgeable and high-calibre human beings. Without them, none of this would have been possible.

This thesis is dedicated to my mother Parvin, my father Mehran, my sister Kimia, my wife Leila and my son Rodean and my grandmother Shahrzad.

Bertrand Russell: "It is not what the man of science believes that distinguishes him, but how and why he believes it. His beliefs are tentative, not dogmatic; they are based on evidence, not on authority or intuition."

Table of contents

List of Figures

List of Tables

List of Abbreviations

BOW Bag-Of-Words
CCG Combinatory Categorical Grammar.
CLTC Cross-Language Text Classification
DT Decision Tree algorithm.
IR Information Retrieval
LDA Latent Dirichlet Allocation
LSA Latent Semantic Analysis
LSI Latent Semantic Indexing
MCMC Markov-Chain Monte Carlo
ML Machine Learning
NLP Natural Language Processing
NMF Non-negative Matrix Factorization
PCA Principal Component Analysis
ROC Receiver Operating Characteristic
SVD Singular Value Decomposition
SVM Support Vector Machines
TF-IDF Term Frequency-Inverse Document Frequency
VSM Vector Space Model

1

Introduction

In literary research, genre theory involves continuous examination of deeper interpretations and complex analysis of textual meaning. Accordingly, Digital Humanities concerns itself with automatic processing of such aspects of literary texts to facilitate literary research as much as possible and to fill in the gaps between literature, linguistics and computation. For example, Ardanuy and Sporleder (2015) used plot structure as a proxy for identifying and distinguishing genre or authorship. In this thesis, we want to automatically extract or get at the semantic properties of the poetic text. The novelty stems from the fact that we deal with Persian poetry, and that we focus on the semantic aspects inherent in such texts. Thus, the context is eastern philosophy, mysticism, Sufism and their rise in Persian history. The poetry classification is a component of a broader architecture of poetry interpretation discussed in this thesis. We present the high-level methodology architecture and the literature behind it and define the research design and questions.

Detecting the time period when a work of art was produced is important. In our work, we propose to do automatic chronological classification of Hafez's poems, into a set of time periods when he might have written them. The main purpose of automatic

chronological classification of Hafez's ghazals[4] is to help us to understand them and also as a consequence, guide the corrections when it is necessary. The objective of this research is to classify the ghazals using machine learning (ML) methods. It is commonly known that any true artist while maintaining his or her authentic individuality, is also more or less affected by the environment they live in. Therefore, the ability to clearly define a specific era and gain a deeper knowledge of the historical-artistic attributes of the time is an excellent source of insights and clues, and these help us understand works of art and reveal the meanings and intentions beneath the surface. For example, imagine if an art researcher could know the characteristics of an era, and how they affected creations such as Homer's Odyssey. The researcher would be far better equipped to analyze, criticize, and understand the work of art in question, and would also gain an advantage since art history is required for true art-related education. Indeed, we can go as far as claiming that the historical essence is an important aspect of any real knowledge. This is particularly true in Hafez's case, as the highly constraining political conditions of the time encouraged a unique type of crypticity and mystical properties to his poems. As a result, ghazal interpretations have been a major source of information for scholarly debate over centuries.

We also hold the work of Dr. Mahmoud Houman in the field of Hafez studies in high esteem. In his book about Hafez, Houman did the chronological classification by hand about 80 years ago (Houman, 1938). Though he did not classify all the poems, it is our understanding that he meant to provide a novel and pragmatic perspective that stresses the use of semantic analysis, as opposed to subjective and intuitive speculation. Thus, our research is essentially a continuation of Dr. Houman's work, using Machine Learning and software automation. We consider this to be a good candidate for ML automatic text classification.

Early on in the process, we began to realize the significant challenges involved. Most daunting was the fact that there was no substantial, reliable electronic Hafez corpus readily available. Nearly all classification tasks in NLP were performed with large corpora, as opposed to our case of 468 ghazals of approximately ten lines each. Very few of the datasets and resources for text classification had poetic content, and none were in Persian. Persian is read right-to-left and uses many dots, oblique strokes and hidden vowels; it is also highly cursive, with many position-driven sub-words. Therefore, we not only had to adapt the software libraries to work with Persian (since researchers have

[4]An average of 10 couplets with the same rhyme at the end of even hemistichs.

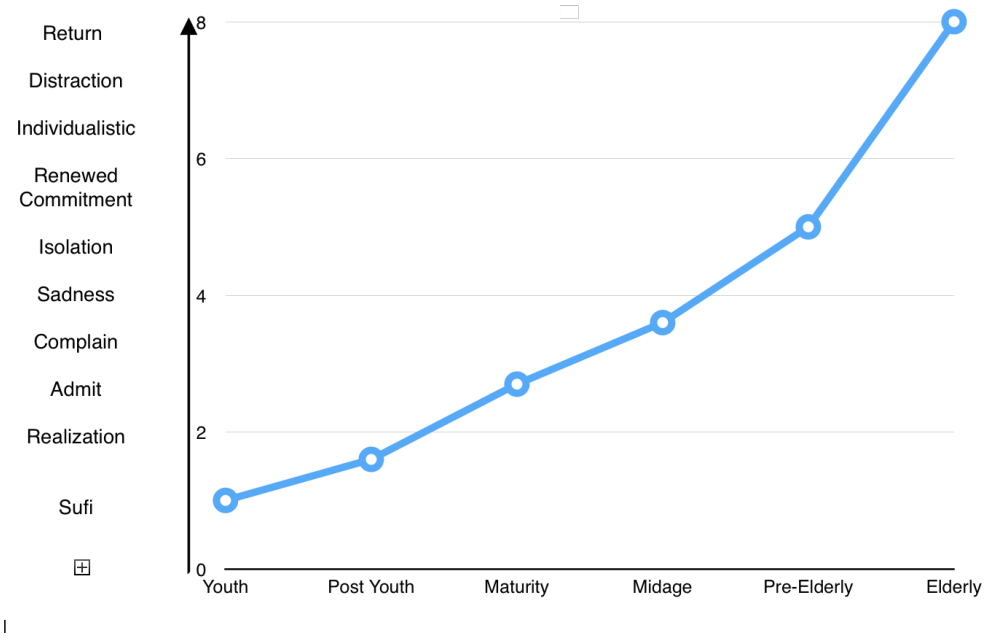


Figure 1: Hafez's Evolutionary Growth Curve

generally designed software libraries to classify English text), we also had to adopt an efficient and accurate classification methodology applicable to our classification task with its small corpus. After we came up with class predictions, an additional but necessary step was to decide how to provide the rationale behind each prediction. Thus, we designed a basic but sufficiently intuitive visualization method for the poems, to address the need to understand the semantic properties of the model and its internal reasoning. A section at the end of the thesis defines some NLP abbreviations we will use in this text.

1.1 Semantic Concepts

Houman provided psychological and personal growth perspectives of the poet Hafez and these play an integral referential role in the interpretation of his poems and their chronological classification. This analytic spectrum of Hafez and his ghazals informed our decision to apply NLP semantic-based methodologies to the chronological classification of his ghazals.

Figure 1 depicts a chronological and conceptual poem chart, with a poem at a specific curve point depending on its determined point in time, based on the semantic elements, themes and attributes which Dr. Houman detected in the poems.

1.2 The Hafez Corpus

We used Ghazvini's[5] version of Hafez, following Houman's approach. We considered consistency as the number one priority during the creation of our Hafez corpora. One of the attributes of an ancient language is its flexibility which provides the freedom to use a variety of writing options in the same compound terms. It should be noted that this flexibility has complex and costly computational implications. We needed to be consistent so that any current or future morphological parsing of the terms is constant across all 468 ghazals. For example, we used multiple types of white spaces to separate one-word terms, and where there is potential confusion, we specified otherwise hidden vowels and diacritics inline.

Our Hafez corpus complies with Houman's order of ghazals; that is, the timing annotation is the actual location of the ghazal in the corpus, with discrete labels. This method was the most efficient means to record Houman's classification, and it set the timing attribute of the poems during the preparation of our Hafez corpus.

1.3 Problem Statement

Our research has three main parts, as shown in Figure 2. Houman's intention for his semantic classification was to support the correction of Hafez manuscripts among 13 different versions. We extend the purpose of chronological classification of Houman to all the ghazals of Hafez using machine learning which corresponds to the second layer of the pyramid in Figure 2. But before any such modelling, we had to prepare the digital version of the poems, which corresponds to the base layer of the pyramid. At the top level of the pyramid from Figure 2, we used the results of automatic classifications to visualize the semantic properties of the chronological segments or classes of poems.

1.3.1 Semantic Modelling, Proposed Methodology

We searched for the most appropriate text classification method for our research and decided on SVM by (Cortes and Vapnik, 1995), which Joachims (1998a) considered a state-of-the-art classification algorithm[6]. Apart from the need for a reliable and consistent corpus, effective SVM feature engineering was also important. There is a huge

[5]Mohammad Ghazvini (1874-1949) was an Iranian scholar who corrected and prepared the most reliable prints of Hafez ghazals.

[6]We used decision-tree and neural-net but SVM performed best.

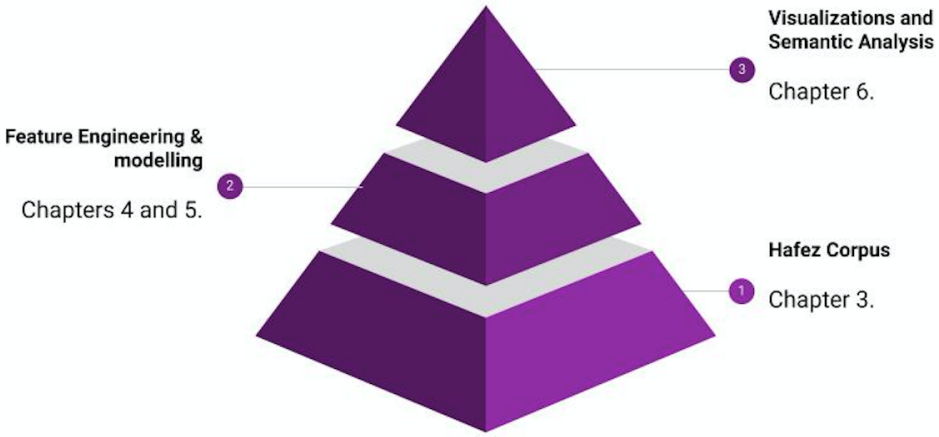


Figure 2: High-level Research Process and Plan

volume of similar work on the application of SVM classification; indeed, there are many related works addressing facial recognition alone (Melišek and Pavlovicová, 2008). LDA and PCA have long been used as effective classification tools in many industrial areas (Marcialis and Roli, 2002) and (Martínez and Kak, 2001). In text classification, many researchers apply LSA or LDA in feature engineering for SVM classifiers (Inkpen and Razavi, 2014). Our feature engineering is based on the layers of BOW, TF-IDF and LDA. As one of our most effective features, we applied our LDA based cosine similarity features to all poems in the training set to determine our top-performing SVM classifier. We used different techniques in isolation, then compared them to identify the best LDA based similarity features for SVM. We have listed the features that proved to be highly effective in chapter 4.

As part of the final analysis of the results, we used PCA to reduce dimensionality and display the LDA results in 2D. The cycle-of-terms, being driven by the LDA model, showed that they bring about and maintain comparable and distinctive characteristics, which help the user distinguish between the ghazals and better justify the classification theme of Houman's classes. As depicted in Figure 3, we grouped the six Houman classes into pairs, and found that the top right cluster has the highest probability terms among the six topics for the class 'Youth', the top left cluster for the class 'Maturity' and the bottom cluster for class Elderly.

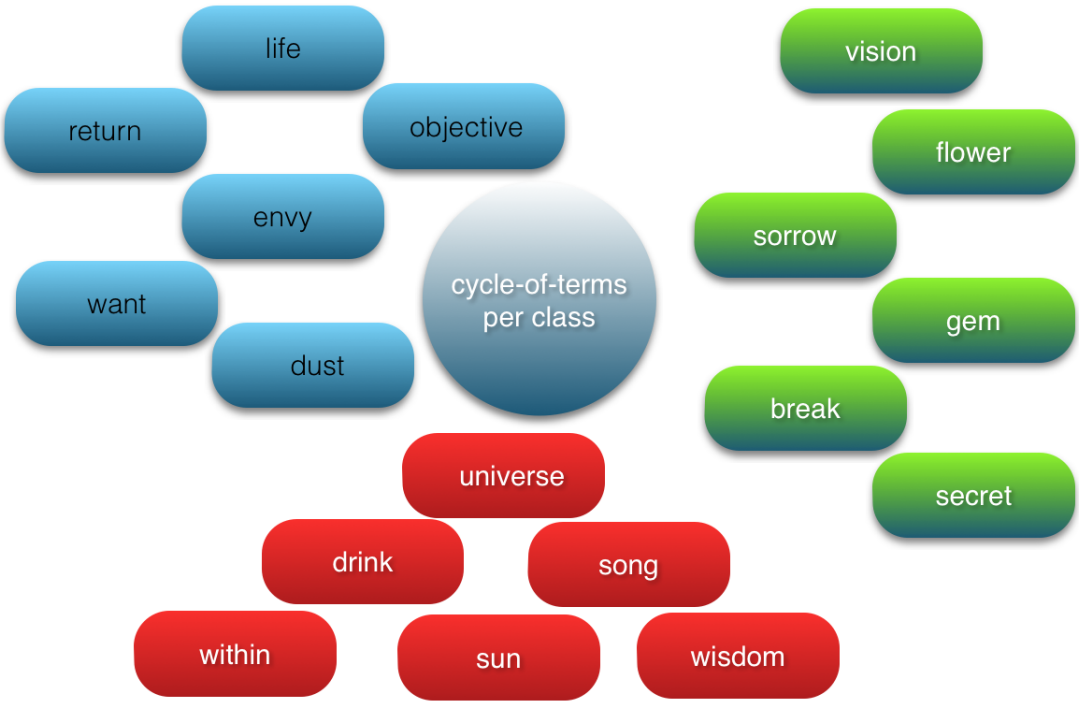


Figure 3: Clusters of words for Houman classes Youth, Maturity and Elderly

1.4 Semantic Modelling, Research Contributions

Although each text classification method we adopted is well-known in the field, the innovative aspect of our research is largely in the feature engineering aspect of the work as well as the application context. To the best of our knowledge, no other layered LDA driven similarity features have been used to create training data for SVM classifiers, particularly in the new application of this method to Persian poetry. Therefore, as a true NLP task in Digital Humanities, there is innovation in working on Persian old poetry with relatively small corpus [7]. The following lists the important aspects of our work:

1. We developed a multilingual Hafez poetry corpus (Persian and English[8]) composed of 468 ghazals, 249 of which are annotated with Hafez classes we can use for training. Of the 249, 21 instances are accompanied by English translations;

2. Hafez Semantic Feature Engineering;

[7]Hafez corpus consists of 468 ghazals, average of 10 couplets each, with approx. vocabulary count of 35,269, of which 14,215 are unique, excluding stop words.

[8]English translations by Shahriar Shahriari are included when available.

3. We successfully created a chronological classifier of Hafez poetry that we used to predict the classification of the remainder of the ghazals that Dr. Houman left without labels; this is published as (Rahgozar and Inkpen, 2016b);

4. We used the highest probability LDA topics as clusters of terms, to represent them as differentiating characteristics of ghazals for their corresponding classes;

5. Hafez Bilingual Corpus Development and Classification; this is published as (Rahgozar and Inkpen, 2016a);

6. We applied PCA analysis to produce visualizations of topics, as well as their word clusters and the relational network for the poems;

7. Clustering of Hafez poems; this is published as (Rahgozar and Inkpen, 2019);

8. Comparison of the two scholars' labels (Houman vs. Mr. Mehran Raad);[9] and their semantic comparisons.

1.5 Organization of the Thesis

The balance of the thesis is organized as follows: Chapter 2 presents the scientific background and related work. Chapter 3 explains the chronological concepts of Hafez poems, historical facts and explains the properties and development specifications of the Hafez digital corpus. Chapter 4 details the classification methodologies we used and describes the classification experiments, as well as the labelling inconsistency management. Chapter 5 clusters the poems, depicts the semantic properties of the clusters and presents an alternative to the two scholarly labelled classifications of Hafez poems. Chapter 6 presents the prediction results and visualizations, compares the semantics of two scholarly labelling, and describes the LDA-driven ontology of Hafez; it also describes the Hafez semantic analysis tool. Finally, chapter 7 discusses our conclusions and future work.

[9]For Prof. Mahmood Houman's and Mr. Mehran Raad's labels, please refer to Section 7.4.

2

Background and Related Work

2.1 What is Natural Language Processing?

Natural Language Processing (NLP) involves automation and processing in applications of human languages. One example in a multilingual context is automatic translation. As the name implies, the objective is to create functionality that allows computers to process human spoken and written languages. The history of NLP goes back to the 1950s when computer scientist Alan Turing developed his ideas of artificial intelligence. But he never wrote explicitly about what we now know as NLP, though Turing test does rely on language processing. The translation is a tangible and complex example of an NLP application. IBM addressed the concept of machines translating one natural language to another in 1954 (Hutchins, 2005). Until recently, the most significant progress in machine translation is due to continuous research projects in the field by industrial giants such as IBM and Google (Li et al., 2014).

An important aspect of NLP is how data is presented to the computer, and much work was done in the 1970s in this area (O'Connor, 2012). For example, chatbot research found that enabling a computer to carry on a conversation and to imitate human dialogue is a difficult task. There are some sophisticated hand-written rule-based systems to search

and determine the appropriate response while collecting and acquiring information from the conversation (Elworthy, 2001).
Regarding methodologies, it was not; 1980's that concepts such as machine learning algorithms were developed to help NLP scale-up rule-based systems and information retrieval methods. Due to the diversity of natural languages and inherent ambiguities, probabilistic methods prove to be an acceptable means of processing language (Manning, 1999).

2.1.1 Applications of NLP

Machine learning (ML) is the field of science that develops algorithms that allow machines to learn from the data so that models can make predictions on the new data and help with decision-making. We usually employ ML algorithms to find patterns in data. However, with the concept of deep learning, a combination of multiple data is used to train the models; representations consist of annotated corpora, dictionaries and hierarchical data.
Once the natural language is involved, one cannot avoid the knowledge representation and inference, which is why strong ties with artificial intelligence were inevitable. Although machine learning cannot completely replace the human ability to interpret poetic text and understand its nuances, it has become an important aspect of the job. Indeed, the systematic analysis of text and its coding scheme, particularly with the high-performance we see today, is only possible with machine learning. See (Alpaydin, 2020) for a comprehensive introduction to ML.
In order to train statistical ML models for NLP tasks we need data (corpora). Unsupervised models can be build directly on the data, while supervised models require that the data is labelled. Before training ML models for classifying texts, features need to be extracted from the texts. Classifiers are trained on the labelled data (training data). Then they can be used to make predictions on new (unseen) texts (that are represented as features, in the same way).
These classifiers need to be rigorously evaluated and measured, and there is an aspect of NLP that focuses on evaluation in order to measure how accurate the predictions were (Sebastiani, 2002). Text categorization or classification is the task of automatically sorting a set of documents into categories (or classes or topics) from a predefined set (Sebastiani, 2002) or assigning portions of text with predefined category labels.

2.2 Text Categorization: Supervised Machine Learning

Hand labelling text is an important process that is used in diverse areas, including literature, marketing research, policy-making, media and researching public opinion. However, manual coding of labelling rules is very costly and time-consuming. So naturally, machine learning has become an attractive option to many scientists in different fields, particularly in NLP to reduce the amount of manual work.
In text classifications as a subset of ML, we employ an inductive process that automatically builds a classifier by learning from a set of pre-classified documents that capture the characteristics of the categories (Sebastiani, 2002).

Text Classification (TC) is a discipline derived from applying ML to text; Manning et al. (1999) dedicated a chapter to it. With text classification, we already have defined categories. Text clustering groups texts by their similarities. Manning et al. (2008) define TC as grouping text item instances.
There are single-label and multi-label tasks in TC (in the latter, an item can belong to any number of categories). A binary TC is a single-label special task that predicts if the document item belongs exclusively to a specific category or its complement. Other related concepts include document-pivoted (DPC) or category-pivoted categorization (CPC). Typically DPC uses a document at the start of the process and then looks for related categories in the document, whereas with CPC a category is the start of the process, and it then looks for all the documents that belong to that class (Kowsari et al., 2019).

In *hard* categorization, a single label is assigned per document, while with a soft approach, there are multiple labels. In addition to categorization with discrete and absolute predictions, there is a ranking categorization that provides a spectrum of continuous results through the document-categories as well. Although the main categorization task is to find a binary relation between documents and categories depending on the requirements and appropriateness, there could be cases that we probabilistically rank by their degree of membership in various categories. Depending on the DPC or CPC condition, we call this process category-ranking TC or a document ranking process, respectively.

The potential applications of TC are extensive; one interesting example is automatic indexing and metadata generation in library systems. Another example is document organization, often used for newspaper articles, ads and patent classification. Text Filtering is another application, in which a stream of documents such as a newsfeed from a news agency is classified according to consumer needs (Hayes and Weinstein, 1990).

The hierarchical categorization of web pages is another area of application (Sebastiani, 2002).

In supervised learning, there is a set of training instances in which the knowledge of the categories is incorporated by manual annotation, which requires dividing the corpus up for training, validation and testing. After the classifier is trained using the annotated training data, its effectiveness is measured and fine-tuned using the annotated validation data. The automatically annotated test data, is used to measure how often the classifier correctly or incorrectly classifies instances. The results are usually depicted by a contingency matrix which shows the predicted vs. actual labels (see Table 1). No test data, in any form, should be part of the training data, otherwise, the evaluation would have no scientific value (Mitchell et al., 1997).

Document Indexing (DI) and Dimensionality Reduction (DR) are fundamentally important, as they determine how texts become meaningful and manageable to the ML classification algorithms. DI and DR should always be applied to the training, validation, and test data sets (Keogh et al., 2001).

One way to represent the text for the classification algorithms is to extract units (features) such as words. This representation is called Bag of Words (BOW) (Zhang et al., 2010). The text is transformed into a vector. The values (weights) in the vectors can be binary (1 if a word is present in a text and 0 if it is not). Another approach for calculating the weights is TF-IDF.
We capture the number of times the term 't' occurs in document 'd' by the term-frequency (TF), while IDF is the inverse document frequency: a measure of how much information the term provides, or how rare it is, For details, see section 4.1.2. TF-IDF is high for terms that are frequent in the document, but rare in the corpus. Terms that are frequent in the corpus do not help the classification, particularly if they appear in many classes (Fautsch and Savoy, 2010).

The indexing method by Fuhr and Buckley (1991) known as Darmstadt Indexing is significant for many reasons, the main one being including the properties of documents and categories and their pairwise relationships in the feature vector. Applications of this method are used to retrieve relevant units of the transformed text based on their probabilistic weights.

From the DR point of view, the reduction can be local or global, and can be done by term selection (e.g. document frequency), or by term extraction where we obtain new terms by combinations or through transformation (e.g. clustering) of the original terms;

		Actual Class	
		C	**Not C**
Predicted Class	**C**	True Positive	False Positive
	Not C	False Negative	True Negative

Table 1: Contingency Matrix

see (Sebastiani, 2002).

We explain the main concepts of other ML methods relevant to our task in more detail in the following sections. Among the supervised methods, we prefer the Support Vector Machines (SVM) approach (Joachims, 1998a). If we apply feature vectors to form decision surfaces, SVM attempts to find the unique surfaces that best separate the classes. We refer to the support vectors of the input as the data vectors of the ML model. Given a set of training examples, with each marked or annotated as belonging to a category, an SVM training algorithm builds a model that can be used to classify other data, and determine if it belongs to a category. SVM is a machine learning algorithm that proved effective in many NLP tasks (Shima et al., 2004).

We typically evaluate text classifiers and their performance and legitimacy empirically, as the central concept of text classification is highly subjective and it would be more difficult to formalize it analytically. Therefore, we usually base the measurements of the frequency of right predictions using test data that is labelled with the expected classes.

We can also average the measurements over all the classes locally, or globally among all documents. We call these micro-averaging and macro-averaging, respectively. Micro and macro averaging are based on true positives, false positives, and false negatives or precision and recall of different data sets respectively. It should be noted that the precision and recall measurements are meaningful individually, but we also consider a measure that combines them, named F-score. Accuracy is the number of correct predictions for all the classes over the total number of text in the test data. Precision is the percentage of retrieved documents that are relevant to a class and recall is the percentage of relevant data that is retrieved for a class. The formulas are presented in Equation 1. F-score in classifications is the equal-weighted harmonic mean of the Precision and Recall measures.

Figure 4: Main Classification Tasks

$$Precision = \frac{tp}{tp + fp}, Recall = \frac{tp}{tp + fn}, F = \frac{precision.recall}{precision + recall} \quad (1)$$

Another evaluation measure is the area under the curve (AUC). By curve we mean the Receiver Operating Characteristic (ROC) curve which shows the true positive rate against the false positive rate at various threshold settings. Using classifier probability membership estimates provides more analytical frameworks for text classification evaluation (Lewis, 1992).

Figure 4 shows the main tasks when classifying text. Due to the general rule in experimental science and the ML methods used in TC, the comparison of classifiers is only possible if we consider three important factors:

1. ensuring the collection of documents and categories are the same among models;
2. applying the same split of training and test data;
3. using the same set of evaluation measures with the same parameters.

2.2.1 Persian Language Classification and Resources

This section describes the NLP resources and tools available for for the classification of Persian texts. The special attributes of Persian from a computational linguistics point of view are notable. Persian belongs to the Indo-Iranian branch of the Indo-European linguistics division. It is actually the Parsi spoken in Iran, 'Arabicized' to Farsi because the phoneme /p/ does not exist in Arabic. We categorize the Persian spoken in Afghanistan, Tajikistan and Uzbekistan as Eastern Persian, which is known as Dari (Seraji et al., 2012). Hafez's poems are written in Parsi.

Persian affects other languages, including Turkish, Armenian, Azerbaijani, Urdu, Pashto and Punjabi, due to the geographic proximities and cultural overlap. The constituent order is relatively free and verbs are inflected for tense and aspect. Verbs also agree with the subject in person and number, but not in gender. Persian lexical items can include what we refer to as 'pseudo-space' or Zero Width None Joiner (ZWNJ), as well as white space in multi-words. The Persian language has both long and short vowels, and the latter is rarely written, only spoken, which causes many homographs in Persian text. Homographs are easier for humans to recognize than for machines (Ghayoomi, 2012).
Seraji et al. (2012) performed an extensive analysis of the orthography, morphology and syntactic structure of the Persian language that impact Persian language resources and their related software tools.
The orthographic properties indicate that Persian has four more letters than the 28 in the Arabic alphabet, but their character encodings are largely the same, except for "ye" and "kaf" which are different in Unicode. Despite popular perception, we propose that there are 33 letters including *Hamze* a*:* which is a glottal stop in the alphabet. Persian does not follow the consonantal root system that is an important property of the Semitic languages, though it is cursive. Cursive means the letterform is a function of its location, such as right, left or dual joining and there are separate codes for variants.

Seraji et al. (2012) states that the phonological and lexical ambiguity is as frequent as in other languages, particularly when the short vowels are left out. We usually skip the diacritic signs, and do not write them in the text. This is a source of ambiguity that humans understand the context.

Seraji et al. (2012) also mentioned the special type of space character other than the white space, that is, the zero-width and non-joiner (ZWNJ). Similar to white space, zero-width space defines word boundaries, while ZWNJ defines boundaries inside a word. For example, *daneʃ* +ZWNJ+*âmuz* is the single word that means "student" and ZWNJ prevents the formation of a ligature.

Persian morphology usually follows an affixable system that has no grammatical concept for gender. Affixes are indicative of pronouns, pluralisms, adjective suffixes and portative genitive, particles that relate with verbs, nouns, and adjectives. Inflectional verbs carry the mood, tense and aspect and agree with the subject in person and number. Apart from the possessive clitic '-e', there are genitive clitics pronouns, as well as plural markers from Arabic that are mainly used for Arabic loanwords.
Persian follows subject-object-verb word-order unless the obligatory subject is embedded

in the verb. Thus, the order of optional constituents in Persian is very flexible, or even omit it. A headword follows its dependent and the verb can be the initial word or the final. The structure is between left-branching (head-final) and right-branching (head-initial) (Stilo, 2005). Unlike the subject, a verb's presence is compulsory (Seraji et al., 2012).

Ghayoomi (2012) used word clustering to create class-based or coarser level terms to improve Persian parsing. He used the modified Stanford parser version for Persian and his PerTreeBank for training while applying word classes instead of words. In this way, the parser can parse any new word if its class exists in the training data, even if the test word itself was not in the original training data. Ghayoomi further showed improvement in clustering by considering POS tags!

The general-purpose corpus of Bijankhan et al. (2011) is the first annotated Persian corpus, and other TreeBank developments often use Bijankhan's work as their base or reference. The corpus consists of 4300 different newspaper articles that are annotated hierarchically by morphosyntactic and partial semantic features.

Dehkhoda (1994) built the main Persian lexicon resource of 343,466 entries with their morphological structure.

Though Persian may possess comparatively less developed annotated resources, there are many valuable contributions in Persian computational linguistics research.

There are also major contributions that apply to most NLP work such as preparing treebanks. The Persian Tree Bank is annotated with HPSG grammar notations (Ghayoomi and Müller, 2011).

The Persian Dependency Treebank consists of dependency relation annotations at the sentence level (Rasooli et al., 2013).

The Uppsala Persian Corpus (UPC) by Seraji et al. (2012) improved the Bijankhan corpus by upgrading or adding the following features:

sentence segmentation, modified POS tags, distinguished multi-word expressions into tokens, defined and tagged pre-nominal clitics, whitespaces changed to pseudo-space or ZWNJ when appropriate, replaced Arabic style Unicode with Persian style encoding and replaced Arabic and Western digits with those in Persian.

Seraji et al. (2012) also created a successor to the Uppsala Persian Dependency Treebank (UPDT), based on the Stanford Typed Dependencies. Their treebank was used in a bootstrapping context to dramatically improve the Persian parser, and the treebank and associated tools are open-source.

Hamshahri (AleAhmad et al., 2009) is another available news article corpus that many researchers and Persian NLP practitioners use.
Another more contemporary corpus is the Persian Linguistic DataBase (PLDB) annotated with pronunciation, grammar, and morphosyntactic properties (Assi, 1997), and there is a bilingual corpus as well (Pilevar et al., 2011). MULTEXT-East POS style annotated corpus of the novel '1984' is also available (QasemiZadeh and Rahimi, 2006). Tools include the Basic LAnguage Resource Kit (BLARK) with its sentence segmenter (Seraji, 2011), a tokenizer, a POS tagger and a parser.
Mosavi Miangah (2009) produced a bilingual English-Persian corpus comprised of web pages and digital documents, to improve concordance in translation.

2.2.2 Cross-Lingual Features and Evaluations

The Cross-Language Text Categorization (CLTC) task involves categorizing text based on the labeled training data from one language, to classify text in another language. The common technique, known as Bag-Of-Words (BOW), can classify texts with up to 90%s accuracy, depending on the task, context and corpora (Zhang et al., 2010). A difference between research works is how they develop and weight features, and they also differ in the learning algorithm they employ to train by using these weighted features. In addition to the BOW features (Nastase and Strapparava, 2013) used etymological common ancestry attributes of words that are shared between two languages (Italian and English in this case) to categorize text more effectively. They also showed that their method improved the baseline classification performance by about a 40% F1 score over the BOW representation. They used the LSA to achieve a better relationship between features or words and document classes. LSA that is based on the training data creates a deeper vector representation of the word-document co-occurrences, through shared lexical and etymological attributes. Rigutini et al. (2005) and Shi et al. (2010) employed source-to-target language modelling based on translation and adaptation. And Wan et al. (2011) and Guo and Xiao (2012) evaluated the cultural and domain differences of training and test data for classification, to highlight cross-linguistic phenomena.
Using LSA is a popular way to create multilingual domain models in a CLTC context. Given the common etymology of the words, Dumais et al. (1997) then applied the SVM classification method to classify mixed languages. The objective is to determine semantic correspondences between languages (Dumais et al., 1997). (Prettenhofer and Stein, 2010) used translation to capture semantically similar words, and to partition the data

with these words into cross-language structures and mappings. Most CLTC techniques use translation and dictionaries (Wan et al., 2011), and induce clusters of words using LSA (Gliozzo and Strapparava, 2006).
In general, the objective of CLTC is to decrease the dimensionality among languages by finding common ground between them. In the literature, researchers have addressed dimensionality reduction using dictionaries. More recently, however, they use the translation with the etymological ancestry equivalents alongside the lexical baseline features. The more the cross-lingual features overlap, the more they transcend language boundaries (Nastase and Strapparava, 2013).

Word etymology is an interesting topic in linguistics and TC and is used to trace back and link the shared words across different languages. Languages adopt words from each other, and to a degree adjust them for different senses while maintaining their common roots. For example, the words “Check”, “Chess” and “Checkmate” in English, have Persian roots of “Shah” and "mât" as in “Shah-mat” (this means "King"-"Mate") and maintain strong semantic relations. A cognate example is the word ‘daughter”, dokhtar in Persian.

Pandian and Karim (2014) addressed authorship identification for both English and Tamil emails, using style-markers apart from the lexical measures, and compared the effectiveness of similar methods between the two languages. They mixed LDA and neural networks for the unsupervised and supervised methods. With regard to feature engineering, they reference authorship identification in other ML works that employ Bayesian regression, SVM, neural networks, k-nearest neighbour and rule learners. They first tokenized and filtered out irrelevant information, then extracted the features based on the work of Khan et al. (2008) and Farkhund et al. (2010). Important types of extracted features, applicable to Persian, include:

1. Lexical features such as the number of words
2. Total characters per line
3. Ratio of digits
4. Ratio of characters
5. Ratio of short words
6. Ratio of uppercase letters

7. Ratio of spaces to total characters
8. Occurrences of characters to total alphabetic characters
9. Occurrences of special characters
10. Lexical word base analysis
11. Number of words, sentence length
12. Average token length, the ratio of short words
13. Ratio of word length-frequency distribution
14. Unique words
15. Syntactic features
16. Occurrences of punctuation
17. Occurrences of function words.

The cross-language aspect of their work is a generalization of their method, which they claim to be language-independent.

2.2.3 Semantic Vectors

Many NLP applications take advantage of 'semantic vector' representations that researchers use to capture the syntactic and semantic properties of the words in a particular language, or in cross-language contexts. Semantic vector representations can be effective in feature engineering of our Hafez classifications as they tend to capture semantics aspects, which are the foundation of Hafez ghazal's and others' classifications. To predict the pivot word, Mikolov et al. (2013) proposed a continuous BOW and skip-gram model in a neural network setting, referred to as word embeddings, which predicts the context by probability maximization conditioned on the pivot word. The expansion of the context window improves the accuracy, but it also slows down the computation.
Pennington et al. (2014) went further, by repressing a word in vector space as a function of a global context by looking at the word-word co-occurrences in the corpus. They then calculated the probability of the appearance of each word in the context of the other word.
Sahlgren (2006) used a Word Space model for semantic vectors, with points in the space

representing the semantic concepts, such as words and documents. Landauer et al. (1997) referred to LSA as a word-document co-occurrence matrix. They then decomposed the matrix into a smaller matrix, using singular value decomposition. Widdows and Ferraro (2008) used random indexing (RI) (Kanerva, 1993) to create semantic vectors and Basile et al. (2009) improved on the matrix factorization and inferential incremental strategy. The random projection concept theorizes that context vectors contain randomly distributed non-zero elements that are assigned to each document. We will use context vectors to analyze documents and create semantic vectors for each term, then calculate the semantic vectors for each document as the sum of all its term semantic vectors.

2.2.4 Word Sense Disambiguation

As mentioned earlier, our Hafez classification task requires capturing semantics. Considering this, it seems logical and predictable to explore an important area in NLP that deals with the meaning of words in context; more specifically, to disambiguate word senses. We apply Word Sense Disambiguation (WSD) applications that are used in translation and anaphora resolution tasks, as the methodology may inspire and help with our Persian poetry case. WSD is a well-known problem in NLP. When a word could potentially have multiple meanings in a sentence or context, we need to decide which sense we used in that context.

Earlier WSD works in the literature use the part-of-speech or other contextual attributes of the neighbouring words in the sentence to determine the best meaning. For example, Navigli (2009) focused on the grammatical and syntactic relations between words and their surroundings. We categorize some research under knowledge-based paradigms such as those referenced in (Lesk, 1986). They use knowledge from lexical resources. They decided the number of senses using WordNet (Miller, 1995), thesauri or special-purpose dictionaries (Stevenson and Soanes, 2003). Montoyo et al. (2005) used both such resources in conjunction with corpus-based or "shallow" paradigms to achieve better word disambiguation. In supervised ML methods, a manually sense-annotated corpus that captures the context of ambiguous words is used to train a WSD classifier. Yarowsky (1994) used decision lists for WSD, by evaluating 100 weighted surrounding words to infer the most probable class of a stemmed word.
This method does not work well for documents with no specific topic; however, Pedersen (2000) applied Naive Bayes ensemble higher accuracy, and Brown et al. (1991) employed bilingual syntactically related words to disambiguate translations with 45%

accuracy. Sarrafzadeh et al. (2011) used Wikipedia articles in English and Persian, and the equivalent of WordNet for Persian called FarsNet for cross-language WSD. Mosavi Miangah (2009) achieved cross-lingual WSD by extracting the most probable senses, counting related word combinations and using the frequencies in conjunction with their co-occurrence in the target Persian language corpus to disambiguate Persian words. Their ambiguous word attributes only included the POS of nouns, pronouns, adjectives and verbs, which restricted and limited translation of English into Persian. Rezapour et al. (2014) used a supervised learning K-Nearest Neighbour (KNN) algorithm for word sense disambiguation in both English and Persian. The Euclidean distance multiplied by weight was the basis for the KNN, as:

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} w_{f_i}(x_{1i} - x_{2i})^2} \tag{2}$$

where w_{f_i} is the weight assigned to the feature f_i. They extracted two sets of features: the set of words that occur frequently and the set of words surrounding the ambiguous words. Their paper brings about a new feature selection process, and a weighting scheme for features. They filtered out and kept the features that led to higher classification accuracy, then the trained a classifier to perform WDS using those vectors as input to the KNN classifier.

In the literature, when we select a sense from a set of predefined possibilities, we refer to the task as WSD supplied with a sense-inventory. However, WSD is a more appropriate term when we divide the usage of a word or differentiate word meanings based on processing unannotated corpora, not only when we employ the semantic vectors discussed in the previous section. SENSE (Semantic N-level Search Engine) uses a combination of both approaches.

The work of Silva and Amancio (2013) is usually in the latter group, and it is interesting that they used a combination of topological or graphical structural patterns inspired by the 'tourist walk' algorithm to capture the semantics and to disambiguate ten polysemous words. This algorithm can conceptualize a tourist visiting locations in a d-dimensional map. The rule is that the tourist goes to the nearest site that he has not visited during the past k steps. The authors achieved good disambiguation performance improvements over the more traditional network characterization measurements and clustering. They represented text as complex networks to disambiguate the ten words used in the context of 18 Gutenberg online repository books. Specifically, they applied the tourist walk algorithm to the graphs with words as nodes connected by edge values to represent the

frequency of words' adjacency in the text. They cleaned the texts by removing articles, stop words and other high-frequency words that convey little or no meaning. The tourist walk algorithm uses the simple deterministic rule that it will visit the nearest node that its control has not visited in the previous k steps; the self-avoiding memory window is k-1. The authors then measured the recurring patterns of connectivity by hierarchical degree, clustering coefficient, average and variability of neighbouring degrees, average shortest path length and betweenness measures. They captured these attributes to train the KNN, a C4.5 Decision Tree, and Bayes classifiers then compared the findings and found that the tourist walk method yielded better discrimination rates than other traditional methods. These are a few examples in which ML algorithms are used to get at semantics, that we also experimented with. Schütze et al. (1995) used mixed models that initially applied unsupervised techniques to WSD by creating clusters and comparing sense pairs. We discuss similar methods in the LDA and Clustering (unsupervised ML) in section 2.3.

2.2.5 Poetry Categorization

There are strong interrelations between lyrics and music, and the properties of one can help classify the other (Baumann et al., 2004). Lyrics analysis also adds to musicology research; from a sociomusicology context, for example (Frith, 1988). Lyrics are often easier to process than audio, and they can play a proxy role for the analysis of musical structure, rhythm and even for melody (Nichols et al., 2009).
Luštrek (2006) did genre detection in text classification, and Simonton (1990) ran experiments in authorship attribution, poetry analysis and lyric-based classification, using shallow features such as POS and function word distribution. Simonton also analyzed the 154 sonnets attributed to William Shakespeare. Each sonnet consists of four consecutive units (three quatrains and a couplet). A computer gauged how the number of words, different words, unique words, primary process imagery and secondary process imagery changed within each sonnet unit, and noticed a common vocabulary change in the couplet. Kim et al. (2010) used deeper features, such as the distribution of syntactic constructs in prose to analyze authorship and writing style. While meter may sacrifice the syntax, rhyme and meter can form effective features to classify lyrics. Scott and Matwin (1998) used synonymy and hyponymy for classification, while Mayer et al. (2008) applied the POS proportion of hapax legomena per document and end-of-line rhyme as features. Hirjee and Brown (2010) developed statistical rhyme detection to

extract in-line and slant rhymes while analyzing Rap lyrics. Fell and Sporleder (2014) classified songs by approximately when the lyrics were published, and detected the genre using features such as vocabulary, style, semantics, orientation toward world and song structure.

Genre classification was accomplished using the Weka implementation of SVM and the following combinations of feature categories (Seyerlehner et al., 2010):

1. n-grams (top 100 n-grams), BOW and collocations as a baseline (vocabulary)
2. POS / Chunk tag distribution as syntactic structure proxy (style)
3. Length (style)
4. Pronouns and past tense verb (orientation)
5. Imagery (semantics)
6. Slang use (vocabulary)
7. Echoism: high relative similarity (style)
8. Rhyme features (style)
9. Use of past tense (orientation)
10. Type-token (vocabulary)
11. Repetitive structure (structure)
12. Chorus (structure)
13. Song title (structure)

Fell and Sporleder (2014) also found that one of the highest performing classifications was for the Rap genre, with its 77.6% F-Score. They showed how lyrics-based statistical models could indirectly help classify music when the lyrics were available. The combination of innovative features used captures the style and semantic aspect of the lyrics, making these features useful in lyric classification.

2.3 Text Clustering: Unsupervised Machine Learning

2.3.1 Latent Dirichlet Allocation

Topic modelling, such as Latent Dirichlet Allocation (LDA) and clustering, are unsupervised learning methods and we often use them as prerequisites to or in conjunction with supervised methods. Unsupervised methods often help enhance the better representation of training data for supervised ML methods. In this section, we examine LDA applications of interest combined with supervised method, and discuss similar mixed and unsupervised clustering methods and their use in literary contexts. We are interested in methods that can perform in situations when the available data is small, as most classification methods require a large volume of documents for training and testing. This creates two challenges, one of which is that it requires these documents to be annotated and encoded as representations acceptable to ML algorithm. The second challenge is the sheer volume of data preparation, which requires human labour. In addition, when considering categorization requirements, we should define the categories (classes) for training purposes. All the supervised ML needs annotated input with set classes. More specifically, the set of features and weight factors for each class or category must form data vectors, thereby allowing us to use them to train classifiers. These limitations encouraged researchers to look for alternatives, particularly in contexts without available large training datasets, and others that are too time consuming and costly or the annotation criteria are unclear. Usual classification methods use BOW, unigrams or n-grams to represent the feature space. Thus, the scarcity of lexical word features associated with each document results in very sparse representing vectors.

Blei et al. (2003) addressed these challenges by introducing the Latent Dirichlet Allocation for document categorization by topic as an unsupervised learning method. LDA is sometimes used in conjunction with other learning algorithms to find more refining features (Inkpen and Razavi, 2014).

In our case, the categorization of Hafez ghazals, scarcity and limited number of ghazals per class, and the labour-intensive annotation of data, compels us to pursue a specific blend of LDA mixed with other ML methodologies, in order to achieve a promising candidate approach. We can also use LDA without other ML methods, as it does not need annotated data to explore and find topics in the texts. Hence, there is methodological overlap with clustering algorithms. According to Blei et al. (2003), the central concept is based on a probabilistic hierarchical Bayesian model that can produce the topics by

induction through the input documents, not necessarily by annotation.

To achieve this, we determine or estimate the distribution of topics over the vocabulary of words using Dirichlet prior, a matrix of words and topics. We also estimate the distribution matrix of documents over topics using Dirichlet prior. This allows us to develop probabilistic distribution relationships between documents and topics, and between topics and word tokens. In this way, words are associated with the topics that we sample for each document. Technically, the advantage of LDA can be attributed to the present conjugacy between the Dirichlet distribution and the multi-nominal likelihood (Inkpen and Razavi, 2014).

LDA assumes the documents have latent topics. Akiva and Koppel (2013) questioned this and presented the topics as top-N highest probability words. Haghighi and Vanderwende (2009) used LDA as a prerequisite to other text classifications. Lau et al. (2012) applied it to word sense detection and Zhao and Xing (2007) used it for translation purposes. Hofmann (1999) applied LDA for probabilistic latent semantic indexing (PLSI). Blei and Lafferty (2005) later found that perplexity plays a counter-intuitive role in topic modelling, and they had people compare the top probable topics to the intruder or outlier words.

Using a 3-point scale, Newman et al. (2010) asked people to rate the topics on how the topic words expressed their observed coherence. They tried to automate the observable coherence to calculate approximations and applied the context sliding-window. They found that the method inspired by Point-wise Mutual Information (PMI) (Church and Hanks, 1990), an alternative to $TF - IDF$ for co-occurring words, is the most reliable method to produce highly consistent document-topic classification. They fed the LDA results to the ML algorithms (SVM, NB and DT) and assisted with the training and testing performance. Employing LDA made feature spaces more manageable by reducing the dimensionality. The results were evaluated using different combinations of SVM, BOW/SVM, LDA/SVM and LDA + BOW with SVM achieving the highest accuracy of 80.4% on FriendFeed data (Celli et al., 2010) and 97.29% in the Reuter R8 data (Aryal et al., 2014).

Blei et al. (2003) used variational inference to estimate the distribution of data vectors, while the work of Minka and Lafferty (2002) was based on the expectation propagation. Griffiths and Steyvers (2004) used the most popular and effective approach that was based on sampling. The Gibbs sampling method was inspired by Markov chains (Porteous et al., 2008), in which the probability of a topic for a word in a document is

conditional on the word itself, previous words in the context and their associated topics. The multiplication of denominators of the conditional probability is the number of times we assigned a word to the topic, as well as the number of times we previously assigned the topic to the document.

The ultimate objective is to make an argument about the probability of using the word (tokens) on the topic and the probability of associating the topics to the documents. Thus, these distributions can create a base for comparisons and optimizations. In other words, the best category has the maximum weight on words-contexts and document-topics that form the best topic candidate for the document.

This scheme is capable of ranking the labels rather than using hard binary classifications. We can extend or tweak the method to help us make the same arguments on multiple labels per document (Blei et al., 2003).

Wang et al. (2007) used LDA for search engine experiments, and Mimno et al. (2011) proposed an alternative method to PMI and log conditional probability to evaluate semantic coherence by highlighting the document-topic consistency compared to human judgments (Newman et al., 2010). The aim of these research efforts is to capture the evaluation of semantic coherence as part of the topic model, which is very promising for our case because it is based on mutual information. In addition, the objective will be to find a cluster of important word meaning in Hafez. Mimno et al. (2011) recorded the collocation frequencies and updated the counts of associated words before and after the Gibbs sampling accordingly for every new topic assignment. They showed that LDA produces more coherent measures than other methods with the log conditional probability. Musat et al. (2011) also attempted to capture the relevance of topics by incorporating the WordNet hierarchy.

Ghayoomi and Momtazi (2014) applied LDA to a corpus extracted from the Persian newspaper Hamshahri and Iran with up to 80.66% and 90.44% accuracy, respectively, and they performed topic modelling and classified news articles using what they called a weakly supervised model. Before they did the final classification they manually shortlisted topics, then mapped the automatically generated topics to a set of predefined topics, which minimizes the amount of human annotation required. Then they obtained classifier predictions by looking up the generated topics, then mapped the corresponding categories to the high-level shortlist. They also used the MALLET toolkit, which is a Java implementation of the LDA algorithm (McCallum, 2002). Ghayoomi's work in Persian text classification is quite recent, and it seems very applicable to our Hafez

classification task as they achieved good results with LDA. Therefore, we believe LDA is a good candidate methodology in the Persian NLP literature, given our limited volume of annotated data. The creative aspect of this work is addressing the numerous topics generated by LDA, and mapping them to high-level news categories such as Economy, Literature and Art, Politics, Sport and Tourism.

Ghayoomi and Momtazi (2014) used the Hamshahri newspaper archive at the University of Tehran, which contains about 318,500 documents. They did not include the annotated labels in the LDA process but used it as the gold standard for evaluation, and to create 35 fine-grained labels. They finally shortlisted the labels and achieved nine coarse-grain categories. When they randomly analyzed the misclassifications they detected errors in the gold standard, which meant that extensive misleading overlap in the standard labels confused evaluation, and the labels had to be verified and corrected. They also found that their LDA model could not assign the correct category labels to the documents in some cases, mainly due to common word-classes. These are aspects and nuances in the corpus and we need to monitor their effects on classification.

Aletras and Stevenson (2013) captured coherence with LDA, which led to distributional semantic similarity between feature vectors for topic words such as cosine similarity and the Dice coefficient (Anuar and Sultan, 2010): $DSC = \frac{2TP}{2TP+FP+FN}$.True-positive (TP) is the number of times the algorithm correctly predicted the class, false positive (FP) is the number of falsely predicted the class, and false-negative (FN) is the number of times falsely predicted that the item did not belong to the class. They used Wikipedia to collect co-occurring words in a five-word window and established semantic vectors of topic-words.

Lau et al. (2013) also used LDA evaluation to capture semantic quality, or discard irrelevant topics. They examined the quality of the topic interpretability by word intrusion (Chang et al., 2009) and observed coherence to emulate human performance; they were more successful using the latter method than the former. They empirically examined different topic-modelling evaluation methods, and proposed an improved formulation of the PMI-based method (Newman et al., 2010). Lau et al. (2013) used the SVM ranking in (Joachims, 2006) to rank topics by word association features and determine intruder word topics. They used PMI, CP1 and CP2 (Lau et al., 2010) in the process and an NPMI normalized point-wise mutual information evaluation measure (Bouma, 2009).

2.3.2 Clustering Methods

As is normal with new methods, some researchers are critical of the LDA approach. For example, the work of Akiva and Koppel (2013) in a continuation of (Koppel et al., 2011) objected to the assumption that each category of interest will necessarily have a distinctive topic, particularly in the context of authorship classification, as each document-author may contain multiple overlapping topics. This is why they criticized the above LDA studies and other work that used LDA for topic-authorship, such as (Rosen-Zvi et al., 2010). The criticisms indicate that they used clustering methods. Graham et al. (2005) decomposed the documents based on their authorship, while Akiva and Koppel (2013) proposed a generic method that required no tagged corpora or training data. They chunked the text into predefined lengths, represented each segment as a binary vector of the 500 most common words in the full text, then measured the similarity of every pair to cluster the chunked pieces into k clusters. They used a precise sampling of purified and labeled chunks as training data input to their SVM method, and claimed to obtain up to 91.5% accuracy for authorship attribution using synonyms as the feature set.
Ghayoomi (2012) used classes of word clusters rather than the word-based clusters, and improved the Persian text parsing precision by approximately 10%. This coarser class level of lexicon based on the similar syntactic behaviour of words has made a significant improvement to Persian parsing using the Stanford Parser (Manning et al., 2014). In this way, if the parser encounters a new word that did not exist in the training dataset, but it recognizes its class, the system could still parse the word according to the class. Data sparsity is a common issue with parsers, and this clustering method helps reduce it. The parser is more genre-independent when using the coarser-level lexicon, and the clustering method helps distinguish the homographs that would have otherwise been treated equally. The clustering assigns homographs to different clusters, according to their POS attributes. This paper used the Brown algorithm as an unsupervised clustering method to create over 700 classes of words. They used the Java-based SRILM toolkit and implementation of the Brown clustering algorithm (Stolcke, 2002).

The Persian treebank is known as PerTreeBank[10] which was derived from Bijankhan[11] and was used in conjunction with the clustering word-class results to train the parser (Bijankhan et al., 2011). This Persian treebank is available online, and as mentioned previously it follows HPSG (Pollard and Sag, 1994).

[10] http://hpsg.fu-berlin.de/~ghayoomi/PTB.html
[11] http://ece.ut.ac.ir/dbrg/bijankhan/

2.4 Visualization and Model-checking

The visualization and model-checking of LDA-based SVM classification have been a very active, open and worthwhile research area (Chaney and Blei, 2012). When semantic attributes and text meaning are searched for, which is one case, often the co-occurrence of terms within documents and the corpus provides good leads, which we deal with technically using probabilistic models. But Houman still demands substantive answers, understandable rationale and detailed explanations. Questions that drive the need for model-checking and visualization include: How many topics should be set as a parameter for an LDA algorithm to optimally obtain semantically distinctive topics? Or how does coarse-graining affect the topics? In addition, given that the LDA algorithm is prior-based, a topic's high probability terms of each run are not necessarily the same. Therefore, another question is how to optimize the number of iterations needed for convergence and to be confident of the significant terms.

The presence of insignificant terms in the topics is an open issue that needs verification by domain experts (Chuang et al., 2012). *Termite* (Chuang et al., 2012) lets the user choose LDA terms from the most probable or salient topic-terms that are domain-specific, but it does not support document-level interactivity or correlations. Snyder et al. (2013) studied the so-called *junk* topics.

Another interesting question concerns stop-words and their effect on classification results. In NLP, stop-words are words that are excluded from training and test data, because they appear in most of the documents. However, in our case classification of poetry stop-words could potentially provide significant benefits, and increase to the accuracy or change the dynamic of topic-terms. Hughes et al. (2012) argued that some seemingly uninterpretable topics could indicate a sub-genre or style.

Topic modelling results usually require expert validation, particularly in industry, thus there is a need for frameworks and tools to assist users to cross-check topic terms and learn the differences between document-topic results in different runs. In other words, we need to facilitate expert verifications and validation processes so they are feasible. Chaney and Blei (2012) employed different document parallel interactivity to observe topic changes. It is usually difficult to understand what is happening inside LDA-based models, which makes providing evidence and intuitive rationale for the results challenging. Such pursuits will enable us to satisfy human curiosity and build trust and credibility, which are the main reasons for developing software tools that can provide detailed system visualizations of internal aspects of topic models. *TopicNets* plots

documents without showing the actual topics using dimension reduction techniques (Gretarsson et al., 2012), and *LDAvis* plots relevance-ranking terms within topics that show their compositions (Sievert and Shirley, 2014b).

Principal Component Analysis (PCA) is a well-known dimension reduction technique that researchers have used throughout the NLP research to find patterns within data. Zhang and Yong Yan (1997) used PCA for facial recognition specifically, which many others then cited also. This led to a body of work on the application of PCA in facial recognition and classification.

There are now many software tools that explore the LDA visualization field, and even a brief description of them could easily take a whole chapter. We use the LDAvis library (Sievert and Shirley, 2014b). One such tool is *RoseRiver*, which utilizes tree and word-clouds (Cui et al., 2014). Another explores topics and highlights their associated terms that relate to documents (Chaney and Blei, 2012). *Overview* (Brehmer et al., 2014) and *VarifocalReader* (Koch et al., 2014) created topic hierarchies based on *TF-IDF*, while *UTOPIAN* utilizes graph and matrix factorization to depict topics (Choo et al., 2013). Given that visualization is not the sole purpose of this research, we will adopt LDA-PCA graphical presentations to assist with interpretation of the results, and discuss the rationale behind each ghazal classification. Refer to section 6.1.3. Our graphical presentation, similar to that of our feature-engineering method is using, among other tools, the *Gensim* library (Řehůřek and Sojka, 2010).

3

Hafez and the Corpus

3.1 Hafez Poems: Chronological Classification

3.1.1 Historical Facts

An understanding of at least a brief history of the era is certainly important to help us get a better understanding of the potential societal effect on the artist and his works. Hafez's lived during 1326-1389 CE according to the Christian calendar[12] (Dehkhoda, 1994).

The second Mongolian attack 1271 CE by Hulagu happened about 30 to 40 years before Hafez's birth. The first destroying attack was in 1219 CE; it is believed that recovery from this attack took 100 years (Dehkhoda, 1994). In Neyshabur even after the Mongolian army left, 400 soldiers were ordered to stay guard to kill any living creature until further notice. Another example indication of the brutality is that the Mongols flooded everywhere in the city and grew hay for their horses. Exceptionally, the governor of Shiraz offered tribute to the Mongols at the time of the first attack, so Shiraz suffered comparably less than other cities.

[12] All dates are converted from Hijri A. H., lunar calendar.

The series of governors of Iranian cities or regions or territories after the Mongol attacks were all Mongols obviously (ɪlχanian). In Shiraz, the Mongol governor was ɑbʊ-æshɑqe-ɱgʊ (1321-1356 CE) who had more of a sybaritic and self-indulgent personality. Hafez seems to have had a job in the palace during his reign and this coincided with Hafez's Youth years: "xoʃ deraxʃid vali dolatæ mostajal bʊd" means "Was joyous but as well short-lived!".

These were very good times for Hafez, given the environment of relative freedom of thought, bars (mei-kadæ) and taverns were open in Shiraz. The next governor is Amir mobarezeddin from Alæ-mozaffar (1318-1357 CE) who was in Yazd then attacked and killed Abʊ-æshɑqe-ɱgʊ and stayed in Shiraz for about a year; during that time all bars and taverns were closed as he was a very hardliner character. This coincides with the third period of Hafez's life. In ghazals, Hafez refers to Amir Mobarezeddin as *Mohtaseb* who bothers everyone and ruins their freedom. In fact, this is to the extent that he puts Hafez in jail and sends him to Yazd for exile (Zendan-e-Sekandar), for the crime of drinking wine!

Shah-Shoja (1357-1384 CE) was the son of Amir Mobarezeddin. Shah-Shoja with the help of his brother Ghotbeddin Shah-Mahmud (1357-1374 CE) blinded his father. Shah-Shoja was the opposite of his father, less fanatic and therefore friendlier; he brought back a bit more freedom and was close to Hafez. Shah-Shoja probably provided him with palace assignments, during this time, or a job at Madresa which was a formal school for religious teachings.

It is said that Hafez was often present in Shah-Shoja's parties and gatherings as more of a musician than as a poet. Hafez played Rubab (a Lute-like musical instrument) and he must have known music very well. In addition, it is said that Hafez had a very good singing voice! Hafez apparently also must have known to play chess and must have known astronomy owing to skillful use of reported terms in his poetry.

There are a few other Alæ-mozaffar governors before Amir Tamerlane, although most of them adopted the Persian culture and language very quickly, even the most liberal ones were highly influenced by the political extremist ideas. Here are the other governors who came to power during Hafez's lifetime:

1. Shah-Yahya (1362-1392 CE)
2. Soltan Zeinol Abedin (1384-1387 CE)
3. Emadeddin Ahmad (1384-1392 CE)

4. Shah MansUr (1388-1392 CE)

During 1369-1404 CE Amir Tamerlane from Turkestan, the other side of Jeyhʊn, another political extremist, attacked Iran and removed the governance of Mozaffarian. He ordered the capture of Hafez and it is said that Tamerlane contested Hafez's reputation. He competed with him and beat Hafez in reading the Quran by memory as Tamerlane could even read the Quran from memory backwards. Nonetheless, he relatively appreciated Hafez's talents and his poetry to the extent that he did not order his persecution!
We see that the volatile environment and the constantly changing conditions of political turmoil affected Hafez and his artworks[13]. He lived during a period of continuous turmoil, change and war. Iran was subject to the governance and attacks of the Mongolians and to their brutality of fanatic rulers and extremists. In a sense, Hafez was a regular being, who was not staying as a pure Sufi or as a mystic throughout his life. His life condition is not similar to that of Mɔlavi's or Attâr's life. However, as a result, his philosophy towards life has gone through a natural maturation process and has been subject to a constant evolutionary change as well, considering the historical aspects and uncertain conditions throughout his lifetime[14].

3.1.2 Hafez Semantics

Khwāja Shams-ud-Dīn Muḥammad Ḥāfeẓ-e Shīrāzīknown by his pen name Hafez (1325-1389) wrote ghazal[15] poems.
His ghazals have been the subject of many interpretations, but mostly the contextual question becomes whether he is criticizing the social settings of his time or he is referring to mystical meanings. The fundamental question is about what Hafez is saying. How do we solve this mystery? Do we have to interpret it as a straightforward day-to-day warning when he advises of hiding the cup of wine because of political threats, or he is in fact trying to teach us about the twists and turns of the mystic subtleties and their metaphors? Are his poems about general life lessons, perhaps referring to traps that man is facing in life and what life is? When he decides to drink wisely as the time is problematic, what are the traces of the influential historical events in his art? What are

[13]There is fiction (no supporting document) that at the time of this attack, his works were burned out of fear by members of his family, which caused his death out of grief. Later on, the works were collected and published by his close friend and admirer Mohammad Golandam some 20 years after Hafez's death.

[14]For more information refer to www.iranicaonline.org/articles/hafez

[15]The structure of a ghazal is very close to that of a sonnet; a poem of about fourteen lines using any of a number of formal rhyme and rhythm schemes; in English typically having ten syllables per line.

the difficulties that he goes through during his life and how much has his era been studied through his lens and how did sociological properties of his time influence his poetry and in what ways?

These questions are important because they help us better understand Hafez's amazing poetry. Is it not possible to assume that Hafez has in fact combined both reality and mysticism in such a beautiful and smooth way that has made them as one unbreakable cohesion? Perhaps there are layers of meaning hidden and interwoven in such an encrypted way together. Who is Hafez? What are the meanings of his Ghazals?

Among many scholars, Ali Hasouri[16] in his book (Hasouri, 2005), following Ostad Zabih Behrouz's[17] perspective which is mostly based on historical events, claims that Hafez is simply a happy poet who is just praising life. Hasouri claims Hafez is really in love with life and his perspective and philosophy towards life is to be happy and to enjoy it as much as we can. Hasouri also claims that if we consider a work of art having mystic attributes and roots, then we should validate the work by the criteria set forward and depicted in the mysticism references in Persian literature. Hafez's ghazals do not qualify to pass those criteria, he claims. He gives examples of concepts and meanings in Hafez's poems that are strongly and directly associated with historical events and are associated with the geographic mentions in Shiraz during Hafez's time. Even if there are mystic references in Hafez's poems, Hasouri advises us to first look through the lens of sociological and historical properties of Hafez's time in order to interpret his poetry.

Secondly, Hasouri says, Hafez cannot be a mystic because he is always questioning the mysticism's fundamentals. Examples are when Hafez refers to the beloved, or to God. If we were to assume him as a mystic, then one wonders why Hafez associated earthy things with God, he asks things from God that do not necessarily fit with the mystic beliefs towards God as the ultimate source. It is as if Hafez does not agree with God's definition in mysticism. Hafez believes in a different God.

Hasouri says that nowhere in the other historical documents about Hafez's time, there is a reference to Hafez as a mystic; but he is mostly referred to as a musician, singer of ghazals and of wisdom. Hasouri follows the suggestion of professor Bastani Parizi that the word Hafez means musician as many other musicians throughout the 300 years before Hafez's time; people referred to those musicians as Hafez. The other secondary but more popular interpretation of the word Hafez is 'the one who knows Quran by heart'; but the musician is the main meaning, Hasouri claims. He also says the knowledge that in fact,

[16]Professor of Persian literature and culture.

[17]Ostad Zabih Behrouz (17 July 1890 to 12 December 1971) was an Iranian playwright and linguist.

a musician is writing these poems is very important in the way we read and interpret Hafez's poems. In addition, Hafez himself does not ever refer to himself as a mystic, Hasouri says. Hasouri has classified only 60 ghazals in chronological order according to their geographic and historical connotations because the other ghazals do not have those elements. For example, if the name of the Shah Mansour is present in the ghazal, then it belongs to that period and so forth.

Hasouri translates the word *Rend* as anti-religion or rouge or knave; but, certainly, Mr. Ashoori, another researcher in this area, has a different perspective in that he believes Hafez is deeply rooted in mysticism (Ashoori, 2009). Ashoori goes as far as to define a new Persian *Rend* of mysticism according to his understanding of Hafez [18].

Darioush Ashoori[19] argues that we should look at any poetry through its intertextuality with other documents; this methodology is inspired by modern social science. Any document has ties and relations with prior books and any other documents of its time. There is a dialogue between these literary works. Hafez is not an exception. One should look at Hafez in conjunction with other works that show Sufi's culture and Sufi's evolution. This discourse has a history. The history of mysticism starts in Shaam and Baghdad. Mysticism comes to Fars and Khorasan and comes into the Persian language. The mystics claim that they knew the ultimate objective of the being, and they got to this understanding through mysticism and by learning and practicing their own religious-driven ways and beliefs.

Mersadol Ebad[20] has indications that Hafez studied and was influenced by Kashfol Asrar[21], by which one can understand many mystic interpretations of existence. *Rend* and *Zahed* are always in quarrel and in arguments. Ashoori asks: What are these coming from? It would be very naive, Mr. Ashoori says (Ashoori, 2009), to think that *Rend* is good but he drinks all the time and *Zahed* is pretentious goodness. So this effect must be only on the surface. In *Asrar* we see that these two archetypes strongly exist and one is referred to as *Rend* and the other as *Zahed*. In *Asrar*, *Zaheds* are the angels that are constantly praying to God. Rend is refers to Adam; Rend is the first man who committed the sin. When Hafez in his poetry calls himself the Rend, he is, in fact, unifying himself with his archetype. We clearly see these parallels in intertextuality. In addition, given

[18]BBC Pargar: Who is Hafez, https://www.youtube.com/watch?v=U8EdY7VLeOU

[19]Born August 2, 1938, in Tehran is a prominent Iranian thinker, author, translator, researcher, and public intellectual. He lives in Paris, France.

[20]This history book was written by Najm al-Din Daya(1177-1256).

[21]A Thirteenth-Century Quranic Commentary, written by Meibodi who lived before Hafez. It is an important book in Persian literature.

that Hafez lived 700 hundred years ago, there is no Newton, Darwin, Freud, etc.; that is, modern science had not come to exist yet; therefore, naturally, the artist's ideology is mythological. However, Sheikh Abu Said is the first positive and a so-called happy mystic that has affected Hafez. According to Ashoori, these are examples that show that we cannot isolate Hafez's ideology from these other giants before him. Ashoori claims that the hermeneutics of what happens to Adam and how God threw him out of Eden has affected the mindset of Hafez. Ashoori says the mysticism based in Shiraz affected Hafez's mystic beliefs.

Shiraz's mysticism brings about very down-to-earth types of ideas to the extent that Hafez most probably believed the place of 'existence' for humankind was here on Earth. This concept is opposed to that promoted by Khorasan's mysticism, which is rooted in much stronger religious-mystic beliefs. The notion of man praising God or the *beloved* has an evolutionary process according to Ashoori; that is, in the early days, the fear created such a belief, the fear between Zahed or the prayer and God. Later, this relationship evolved into a loving one that was based on the passion between the lover and the beloved. This concept is consistent with the intertextuality of other documents, Ashoori claims. To Hafez, God becomes poetic and beautiful and a kind being from what it was before, a revenging and scary character.

Hasouri, on the other hand, disagrees that the meaning of Rend is necessarily consistent with this mystic evolutionary process and hence disagrees with the new perspective of loving God in Hafez's ghazals, especially without any substantial evidence to supporting it.

Ashoori says that the deeper analysis of the semantics of ghazals indicates that religious locations such as mosques are portrayed as negative places of wine drinking which is praised with positive connotations. This concept corresponds with the dual archetypes present in the intertextuality mythologies of Hafez's era. Hasouri does not necessarily deny the presence of these semantic elements but disagrees with the mystic arguments and interpretations. Hasouri says Hafez uses these terms and symbols for his mostly anti-religious objectives. *Rendy* (verb) means "being *Rend*" (name), to Hasouri is anti-religious, but to Ashoori is a reference to Adam and Adam's rather unfair destiny and his hardships.

Apart from all the nuances of either realistic or mystic perspectives in the interpretation of Hafez's ghazals, Ashoori looks through semantics, with sociological lenses. Ashoori studies the intertextuality of the evolution of words semantics and therefore goes about

removing layers of meaning to better understand the words' connotations. The words find a different sense in the context of mysticism and hermeneutics. For example, behind the surface of the chaperone meaning of the word *raqib* Ashoori finds references to the competing relations between Adam and the angels.

Mahmoud Houman (Houman, 1938) presents Hafez as a very special poet whose poetry includes mysticism, resentment against pretentious clergy, demeaning towards dialectic philosophy, while praising love and *Rend* aspects. Houman strongly sees and encourages the deep interrelation of the ghazal lines and, as many others claim, the first line often carries the main topic. He promises that understanding Hafez is, in fact, possible; and although there is no supporting documentation available to us other than his poems, one can still understand them by setting aside personal tastes and presumptions and then only using logical analysis.

Houman categorizes Hafez's thoughts into main and secondary concepts by studying genealogy and the semantics of the symbols and expressions used in the poems. Some of the main concepts are 'knave'[22], 'love', 'wine'; some Hafez's secondary concepts are 'destiny' and 'dervishi'[23]. Through this logical analysis combined with the supporting geopolitical and historical events at the time of Hafez, plus his psychological properties and his philosophy, Houman is able to map Hafez's poetry to the evolution of his thinking and accordingly maps his poems into different time slots of Hafez's life.

Houman believes in Hafez also as a philosopher-poet who questions the universe and who does not seem to believe in the plots set forward by the religious documents. Houman brings many solid references from the ghazals that support his claims. For example, the last line of a ghazal reads: 'Hafez, our existence is a mystery, solving it is but all imaginary'.

Houman's classification of the ghazals is based on the intrinsic evolution of the poet and therefore he analyzes this evolutionary process starting from questioning Hafez's philosophy of being and the way Hafez views the origin of the universe, as a result, Houman is able to explain why Hafez is astonished.

Houman follows the semantics of expressions through the prior literature and cultural history, and draws such conclusions that relate the perceived "wine drinking" with demeaning intentions towards clergy and religious ideologies; he brings examples of other

[22]English Dictionary translates Rend as a dishonest or unscrupulous man.

[23]Google: A member of a Muslim (specifically Sufi) religious order who has taken vows of poverty and austerity. Dervishes first appeared in the 12th century; they were noted for their wild or ecstatic rituals and were known as dancing, whirling, or howling dervishes according to the practice of their order.

giants of Persian literature such as Rumi and Attar who use similar expressions to bring down the dialectic and logic. Houman interprets the drinking wine and its praise by Hafez as a symbol of dislike towards the blinded science that is solely based on logic and has no legitimate philosophical context.

Houman also does very interesting psychoanalysis on the bipolar emotional reasoning behind Hafez's preferences; for example, Houman mentions love against knowledge, praising the unworthiness of the world against notions of power and affluence. Houman presents Hafez's attraction toward turning his back on social rank by praising doubt. Houman mentions the praising love against imaginary hopes and religion are all worthwhile clues and indications.

The meaning of 'Rend' that is 'knave' in English is somebody who is against the asceticism and the ascetic. Examples of such uses are brought forward in the works of Khayam and Sanaii to support this use of "knave". Hafez's behaviour by turning back to affluence and materialism, demonstrated in his ghazals, according to Houman, is rooted in Hafez's perception that pretences and charlatanism come from the weakness of character. Weakness, in turn, is a roadblock to human greatness of self; hence, one should avoid attractive superficial and decorative earthly materials and instead thrive for simplicity, and strive and search for cleanliness in psyche and constantly endeavour for truth.

Hafez also has three types of references to love in his poetry, according to Houman: Freudian, Platonic, and mystic love. These types are contradictory and therefore cannot happen at the same time but they exist throughout the life of Hafez and throughout his poetry. Hafez never stayed in one type according to Houman but he even had the knave's attitude towards love itself. Houman brings much supportive evidence from the ghazals where Hafez shows love towards righteousness or God, and refers to humans as a realization of righteousness. Wine also closely relates to the meaning of love; wine essentially makes love come to life.

Hafez has references to destiny as cause and effect, and sometimes as an unchangeable and predefined event. Houman maps these philosophic changes to his maturity growth line. Similarly, he has references to dervish attributes to reference the freedom and sometimes to mean withdrawal and contentment. Houman uses these concepts and analysis results to map the poems to the chronological evolution of Hafez's character and ideology and his maturity growth.

3.1.3 Divan of Hafez

Houman used the compilation of Hafez's Divan by Mohammad Ghazvini and Ghasem Ghani (1942). Erkinov (2002) used earlier copies of Hafez such as the ones by Khalkhali and Pezhman (Arberry, 2004), each had used 6 to 8 different older copies of Hafez to compile, develop and correct their copy. Therefore, Houman claims that indirectly his final copy inherits from about 16 copies of the Divan of Hafez. Houman's classification of ghazals was initiated by the observation of differences in the ghazal's copies. Houman found Ghazvini's corrections subjective so he observed a gap and necessity for a more systematic framework to guide the correction of Hafez poems. Houman (1938) counted that majority (70%) of ghazals diction were common, reliable and correct across different copies of the Divan, he hypothesized that the 30% could be corrected based on the larger segment of the Divan. He decided that two aspects of ghazals could form more reliable criteria for the correction, validation or decision over the correct form of the words in the ghazals that had few differences in the copies of Hafez. One aspect was semantic and the other was formal aesthetic or style. The former is the subject of our research, the latter is outside the scope of our current research and left for the future.
Hafez uses multiple types of poem structures. There are 468 *Ghazal*, 2 *Masnavi*, 3 *Ghasideh*, 30 *Ghatae* and 40 *Robā'ī*. These short and long structures are known for a specific theme, style and purpose. Houman has classified only a subset of the *ghazals* Houman divided the ghazals into six main classes, which correspond with the approximate segments of Hafez's life chronologically. In Houman's mind, each class contains only the poems that are similar in context and follow Hafez's specific worldview during that segment of his lifetime. See Table 2.

Class	**Youth**	**PostYouth**	**Maturity**	**MidAge**	**Before Elderly**	**Elderly**
Count	38	25	79	66	28	13

Table 2: Lifetime Ghazal Periods

Each ghazal has an average of 10 pairs of hemistichs. The first pair or the onset usually has the main theme. The last pair has the Hafez name, in which the poet refers to a pseudonym or enunciates himself. Houman only classified 249 Ghazals and 219 were not classified. In his Hafez book page 377, Houman (1938) says that he had only finished 249 of the ghazals classified and the rest were work-in-progress but he mentioned that

they were sufficient to demonstrate what he had set forward to achieve. Houman make his points about semantics and style of Hafez's ghazals and the logic behind their classifications. Houman provided an objective framework and guidance for interpretations and corrections of Hafez poems. As an example of the correction of Hafez's poetry, in Ghazvini's version, it is written "thank god ..." whereas the correct form should be "thanks to..." (Houman, 1938: page 417, poem no. 184 - 238, correction 66). This poem belongs to Elderly or Senectitude or the very last class of ghazals classified by Houman: *In our midst,* ***thank God***[24]*, the dogs of war are put in chain and lock* ***The angels*** *gratefully drink, gracefully dance, from block to block.*

According to Houman, the parallel between the two hemistichs would be more intact with this pattern of "thank that..", which is also present in other undisputed ghazals. Also, it is possible to replace "that" with "god" in Persian hand-writing. Therefore the Khalkhali version of Divan-e-Hafez must be the correct version. Also in Ghazvini's version, the word "angels" is written "Sufis". This cannot be correct and angels should be the correct version consistent with Brockhaus's old version of Divan-e-Hafez (Brockhaus, 1875). According to the proper classification and meaning of the poem, there is no reason for Sufis to celebrate the partnership of Hafez with his beloved. In Houman's ontology of Hafez, Sufism more or less equates with sadness and obstinacy. Sufism ideals contradict joy, celebration and drinking. On the contrary, angels' celebration is not only quite consistent with Hafez's philosophy and semantics of his ghazals but is also conforms to his aesthetic and style, showing the grandeur and depth of his spiritual love during Hafez's senectitude.

Another example of Divan-e-Hafez's correction by Houman (1938: page 418, poem 355 - 248, correction 67) is from old age class or senectitude:

If I am the rascal of the tavern or the ***Hafez*** *of the city, I am that which you see or even less*

In Ghazvini's version, it is Zahid or Sheikh or the preacher. According to Houman, Hafez never associate or calls himself the sheikh or the preacher. In fact, he belittled or resented them during the second and third period of his life. It would be therefore very unlikely that he ever called himself that, let alone doing that during senectitude. This correction is consistent with Brockhaus' old version of Divan-e-Hafez which is also consistent with Hafez's style.

[24]Shahriari's translation, intentionally or inadvertently, is following Ghazvini's version.

3.2 Hafez Corpus

In this section, we introduce many aspects of the Persian language that could have implications in natural language processing applications. Although there is incredible work done in the preparation of Persian corpora such as Bijankhan et al. (2011) and AleAhmad et al. (2009) and there are many valuable NLP works for Persian (Seraji et al., 2012), we believe that the research community has only scratched the surface. We see much room to do justice to the digitization of Persian text. We need to apply defined linguistic rules consistently and clearly throughout the text to make it ready for NLP applications. In this section, we try to make further progress on that front while preparing Hafez corpus. In the computational world, the Persian alphabet is an extension of the Arabic letters plus four extra. Apart from a few disparities such as for kâf and *"ye"*, characters are represented in almost the same Unicode UTF-8 in both languages depending on the software system and adoption policies. We need to consider the overlap of common characters in preparation of Persian text for automatic processing and we should define and account for such linguistic aspects. We introduce the Persian language properties and the implementation aspects for preparing our corpus. We specify how we have overcome some of the ambiguities in the following sections.

We obtained the poems from Houman's Hafez book that Mr. Esmail Khoi edited. We typed them in electronic form while following the wording and order from the book. In addition, we have made sure that the words are typed consistently both from lexical and grammatical perspectives. We have tried to follow the guidelines of the Persian Language Academy of Iran wherever possible.

We look at the linguistic aspects and properties of Persian such as its orthography and morphology that are applicable to the creation of the Hafez corpus from the information retrieval point of views. These aspects will provide requirements. Central to all is that we have applied three types of spaces and we have used them to overcome word ambiguities. Especially, we emphasize on the use of the third type of space as novel.

3.2.1 Persian Orthography

The cursive script nature of Persian means that the same character may be written in linked or stand-alone forms depending on its location in the word: initial, medial, final. The form is also a function of the adjacent letters in the word in question. There are Dual-joining and Right-joining characters (Seraji et al., 2012)

Because of the variety of possibilities, it is easy to sacrifice consistency especially in writing and preparing a corpus for computational processing. In classification, consistencies in orthography are of the essence otherwise the same intended word would be introduced as different and consequently affect the classification accuracy.

The purpose of white space in Persian is not just defining the boundaries of words, because of the cursive nature and other orthographic aspects, space also plays important roles within the words. Therefore, the other two spaces are non-printing spaces. We refer to the second type of space as pseudo-space or zero-width non-joiner (Seraji et al., 2012). Take, for example, in the word for student in Persian, between "danesh" and "âmuz" there must be a pseudo-space to make them as a single word, although we write it as a non-linked cursive. Inflectional suffixes are examples of lexical ambiguities that pseudo-space resolves.

We refer to the third type of space as *joiner-no-width-space*; its role is the opposite of the second type of space. We use it to differentiate and separate the linked cursive subjects and plural markers mostly linked to nouns and verbs. For example, didamaʃ means "I saw her." , which needs a *joiner-space* between "am" *I* and "aʃ" *her* so that the system can distinguish and separate the subject and the pronoun, both of which are linked to the verb.

3.2.2 Persian Morphology

According to the guidelines of the Persian Language Academy of Iran, we should write the comparative adjective made by postfix "tar" as separate. For example for the word, "kuchak-tar" which means "smaller", it is possible to write the two parts separately. However, with some words such as "keh-tar" and "meh-tar", this guideline makes the corpus ambiguous. "bish-tar" means "more" and we are supposed to write it separately not linked. Therefore, in such instances, we have linked the postfix. However, we have tried to be consistent as opposed to many available Persian electronic documents that write the same words in different ways.

Another example is the plural postfix *hâ*. Persian Academy suggests that we write it separately. For example, we had to decide, do we write the word animals as separate "heyvan-ha" or linked as "heyvânhâ". We chose to comply as much as possible and, more importantly, be consistent across the corpus.

Following the guidelines requires multiple types of spaces. Apart from the regular white space, we have also used two other; of the pseudo space and the no-width optional break.

For example, we used them for the plural postfix "hâ" wherever it did not link with the word to make it as one word. Another example, for the third type of space, is with the "mibinamet", which means, "I see you", but is written as linked.
We categorize the morphological implementation rules that we have considered and applied to the corpus as follows:

1. Possessiveness or ezâfe is written as genitive clitic or pronominal genitive clitic.
2. Plural markers differentiation for Persian words and for those borrowed from Arabic.
3. Differentiation of the comparative and superlative adjective suffixes.
4. Pronominal clitics attached to nouns, verbs and adjectives.
5. Pronominal clitics attached to adjectives, adverbs, prepositions, verbs.
6. Non-canonical subject.

We have not defined any of the verb-subject rules in our annotations. We could do this in the future works. We conjecture that this could increase and improve classification accuracy when using POS features. This means that the current classifier does not specifically separate to distinguish the links between subjects and verbs and so we consider such linked entities as single terms. For example "mibin-am-et" is written as one word but should be recognized as three: "I see you".

3.2.3 Corpus preparation: Summary

We summarize the implementation specification of our Hafez corpus development as follows:

1. Consistency in writing when the word can be written in different ways.
2. The proper use of glottal stop or plosive consonant *hamze (aː)* whenever appropriate.
3. The proper and inline use of diacritics whenever it was necessary.

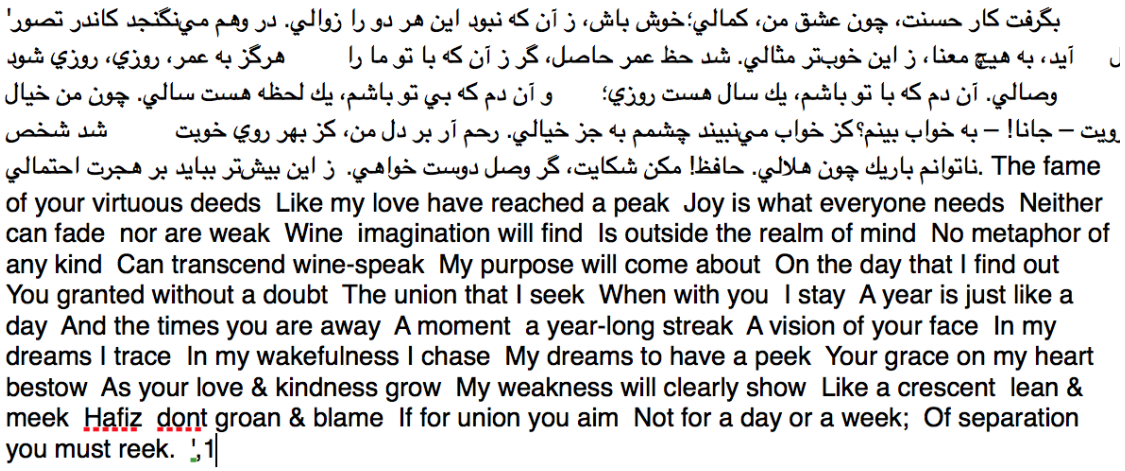


Figure 5: Hafez's poem in digital encoding

4. Application of second and third type of white-space to foresee the support for the future efficient POS annotation and parsing.

5. Following the many diction rules that are derived from the feasible definition of Persian parsing requirements.

The corpus refinements are not by any means final but we are on the right track. For more details, please refer to the appendix: Some Considerations for the Persian Language. Figure 5 is a snapshot of the digital encoding of our Hafez corpus; we see a ghazal in Persian accompanied by its English translation; it is labelled 1, as class Youth, for training purposes.

4

Chronological Classification Methodology and Experimentations

4.1 Classification Method

In this chapter, we describe the method we used for classification, including two main challenges. One issue was that our classification drew on a relatively limited number of instances. The differences between poem classes or categories are very subtle and fuzzy, and one cannot distinguish them without a deep analysis of a poem. Hafez's ghazals are the product of a well-educated, deep-minded and high-caliber artist, who did not drastically change his philosophy and perspective on existence during his lifetime. It is safe to say that, as with other artistic work, traces are not direct, and through time there are changes in innovative perspectives and evolutionary ideas within multiple layers of meaning. The other issue was due to the limited number of ghazals in our corpus (468 ghazals), which is considered very low in the text classification field. Thus, finding a suitable technique with acceptable performance and accuracy to serve our purpose was very difficult.

A combination of ML and rule-based methods typically performs best in real-world in-

dustrial applications, particularly when new business imperatives require quick response implementations to manipulate the data, something that rule-based methods do very well (Villena-Román et al., 2011). Though the ML methodology is our focus, in practice one can couple our techniques with any other complementary rule-based implementation or technology. As stated by Manning et al. (2008), this strategy is the main reason why decision tree type ML methods are so popular; they provide more interpretable classifiers since they encode rules in tree format.

Our chronological classification of Hafez's poems involved a combination of techniques to improve performance. The selection of techniques is based on the most effective semantic-based text classification methods in the literature and, empirically, they have been quite effective in our case. SVM (Cortes and Vapnik, 1995) proved to be a state-of-the-art classification algorithm for applications of topic-based text classification such as ours (Joachims, 1998a). Colas and Brazdil (2006) compared SVM with other algorithms. However, we have adopted a multi-stage approach that employs topic modelling techniques in conjunction with SVM algorithms, including LSI, LDA and PCA, as the feature engineering techniques to transform and prepare the training data for SVM. Thus, to improve performance we used LSI and LDA as prerequisite steps in our feature engineering and preparation of training data before it is input to SVM. We use a BOW representation with TF-IDF weights before applying the LSI and LDA unsupervised models. We also used an LDA-PCA model independent of the classification, for visualization, discussion and further analysis of the results.

Our Hafez corpus is labelled according to Houman's classification in which 249 of the 468 ghazals were classified As mentioned, the Hafez corpus is very small but what matters most is to achieve acceptable accuracy and meaningful evaluation results on our classification models. The priority was to predict the classes for the unclassified poems according to the labels defined by Houman, and our top model satisfied these criteria for the six classes. However, at high levels, as we show the experiment in Table 6 in section 4.1.11, we did a two-phase classification with a reduced number of classes. Our strategy was to amalgamate in order to reduce the number of classes and increase performance, and this approach was consistent with the literature guidelines for SVM, as proven by our results.[25].

We presented the option of combining the original classes into two or three while maintaining the chronological order of the ghazals defined by Houman's classification. That

[25] Experimenting with BOW features for the SVM classifier for both cases; as expected, the smaller number of classes had better performance.

is, in any of the SVM models, the number of classes will be either six, three or two. For example, in the first phase, we train and predict with three of Houman's amalgamated classes. The first is the Youth and Post Youth class, the second is the Maturity and Middle age class, and the third is the Before-Elderly and Elderly class (Rahgozar and Inkpen, 2016b).

In all our Hafez experimental designs, we maintained the chronological order of the labelled classes with respect to Houman's well-defined order classification; in the second phase, we drill down into the next level of granularity. For example, if the model prediction is classifying an unclassified ghazal to our second amalgamated class, in the second phase we attempt to predict and classify the ghazal to the Maturity and Mid-Age class. These are a more fine-grained breakdown of our assumed class with regard to how it was originally defined and created. Ultimately, we accurately classify ghazals to whichever of Houman's six classes they belong.

The base methodology is the same in all experiments. We use SVM in the second stage, while the training data features of BOW, TF-IDF, LSI or LDA-driven term distributions are prepared in the first stage. Also, in some experiments, we used the similarity factors based on the LDA model. In this chapter, we explain each technique individually.[26].

We used SMO, the multi-class version of SVM that is implemented in JAVA by Weka (Hall et al., 2009). And for feature engineering and graphics we used the open-source Python package genism (Řehůřek and Sojka, 2010).

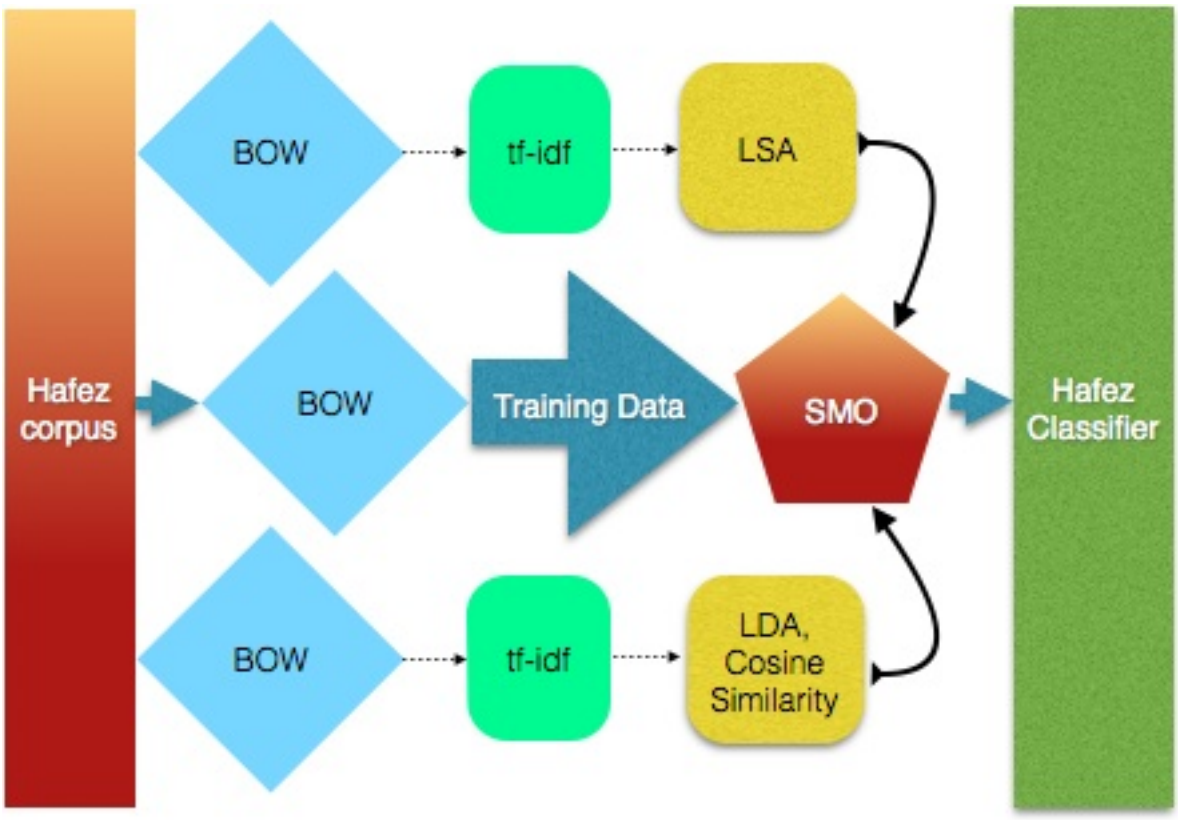
We explain the techniques we applied in this section, and provide applications and references from the literature.

4.1.1 The Main Modelling Components

From the start of our classification experimentation, we pursued a variety of feature engineering endeavours to improve performance and finally determined that LSA and LDA-based features are highly effective classification techniques for Hafez poems. We are not aware of any work in the literature that specifically employs LSI or LDA Cosine Similarity features to represent the training data for SVM to classify poetry with such a small corpus. However, Inkpen and Razavi (2014) did use LDA and SVM. Kwok (1999) discussed mathematical interrelation with Bayesian models to moderate SVM output, and explains the relationships between the evidence framework and SVM. Conversely,

[26] We used classification algorithms other than SVM, including NB and KNN from Weka, which produced less accurate results.

Figure 6: Main Classification Methodology



Shima et al. (2004) used SVM output to improve LSI classification in multi-stage processing.

As shown in Figure 6, there is a sequence of transformations of the corpus . We created the Bag-of-Words (BOW) first and used it to calculate the TF-IDF, as well as the dictionary index for the entire corpus. We refer to our approach as two-fold feature engineering and using the TF-IDF we created the LSI measures and LDA similarity measures. We included three feature types in the final training data for SVM: the BOW, the LSI and the LDA similarities. As mentioned, we used the Weka (Hall et al., 2009) multi-class version of SVM, known as SMO, and the LSI and LDA Python implementations from Gensim.

A description of the techniques used in the modelling steps follows.

4.1.2 BOW and TF-IDF

BOW and TF-IDF are likely the most fundamental features for training ML algorithms for text classification. We used the BOW features for SVM as our first experiment and then endeavoured to improve the results by employing other techniques. BOW is typically a very strong and effective feature set for text classifications, and as such there have been attempts to build other innovative features on top of it. BOW involves the preparation of the list and counts of words in the corpus. The word counts are usually normalized, and the word count matrix for each document makes the sum of counts one per document. We normalize by dividing by the length of the document, which coincides with the axiomatic fundamental definition of probability. The normalized frequencies of words are essentially the probabilities of terms within documents.
TF-IDF attempts to go further, by providing weights according to the specific document relevance of the term, as opposed to just its frequency. We determine the relevancy of a term in the document by dividing its frequency by the overall frequency of documents in the corpus that contain the term. Thus, we calculate every term *TF-IDF* as follows:

$$W_{t,d} = tf_{t,d}.log\frac{\mid D \mid}{\mid \{d' \in D \mid t \in d'\} \mid}$$

$tf_{t,d}$ is the frequency of term t in document d. $\mid D \mid$ is the total number of documents in the corpus. The denominator is the total number of documents that contain instances of term t.
We calculate the *TF-IDF* factors for each term of the document and index them with a dictionary or a hash table so that we can combine the features correctly and so they can be identified as elements of a document vector.

4.1.3 SVM

Much research on the SVM and its application in diverse fields has been conducted over the past 20 years, as well as other equally important state-of-the-art ML applications using decision trees, regression, neural networks and random forests. Only NN methods are the subject of new research. Our objective is to identify applicable aspects of SVM, evaluate their engineering parameters and generally explain the important sections of the algorithm, particularly the engineering and development properties. We focus here on the engineering aspect in the text classification algorithms, in the role of ML practitioner rather than ML theorist. This approach is especially important, because not every classifier performs well in every application, or with specific types of data sets.

The lack of labelled data was not an issue for us, as we were able to prepare it ourselves. Extracting the rules would have been a significant separate undertaking on its own since the volume of our data was also relatively small.

ML theory also recommends the use of methods such as Naïve Bayes, that apparently perform well in these conditions, as reported by (Ng and Jordan, 2002) and (Forman and Cohen, 2004). However, there is some doubt and controversy about NB performance (Manning et al., 2008) when it is applied to text documents. SVM is usually more accurate than NB, but the trade-off is that SVM is slower with large data sets. Our data is very small. The literature does not advise using models such as the nearest neighbour with a limited amount of data. In our case, we achieved excellent empirical accuracy using SVM, as shown in the experiment discussed in section 4.1.10. The unsupervised feature engineering methodology we employed was perfect for SVM.

There are many versions of SVM that are important and relevant to our purpose; (Joachims, 1998b) and (Platt, 1998) developed refinements on the SVM algorithm and used it in text classification. SVMs are inherently two-class classifiers, while the multi-class versions are often based on the "One-versus-All" technique, to utilize the binary behaviour and extend it to multi-class through the iterative combination of binary classifiers. A more elegant version of multi-class SVM is based on the maximization of the weighted feature vectors of the pairs. If the class data sets are not linearly separable, we call the SVM model nonlinear. The predicted probabilities are usually paired using Hastie and Tibshirani's pairwise coupling method (Manning et al., 2008), (Hastie and Tibshirani, 1998).

As shown in Figure 7, sourced from (Manning et al., 2008), the data could be linearly separable by the hyper-planes, in our case. As we had six classes, we had to use a multi-class implementation of SVM. It was intuitively feasible to conceptualize and visualize our six classes. The difference between many ML methods is how they function, or the criteria they draw the separator hyperplanes from to form the decision boundaries between the classes. Perceptron algorithms find separators, Naive Bayes uses probabilities conditioned on specific criteria and SVM finds the hyperplane that is farthest away from the support vectors and the points on the margin, as shown in Figure 7. We call the distance between the decision hyperplane and the closest data point the margin of the classifier. The chosen support vectors are the set of data points that are qualified to form and participate in the constraint maximization system. The solution to the objective or decision function provides the location of the separator, constrained by the support vec-

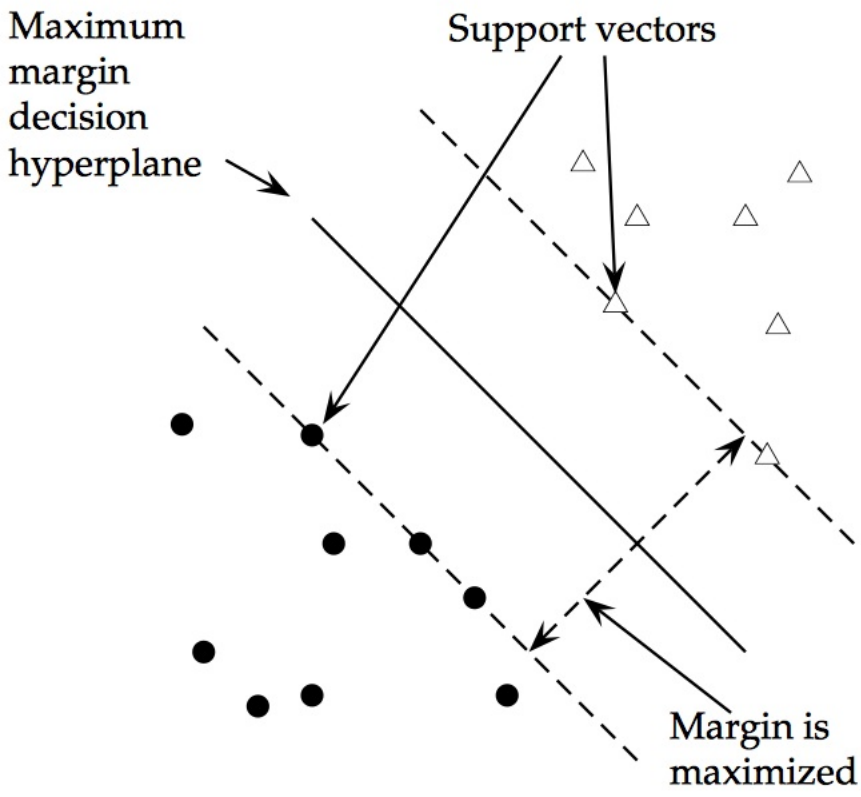

Maximum
margin
decision
hyperplane
Support vectors
Margin is
maximized

Figure 7: SVM

tors.

The following paragraph, which deals with ML technicalities, may be omitted without break in continuity. Denoting the feature space vector by transformation z, which we calculate using the kernel function $z = \phi(x) = k(x, \hat{x})$, delivers a similarity measure to help us find the separating hyperplane, with x as the training data and $\hat{x}$ the unlabelled data. In addition, if α_i are Lagrange multipliers for input pattern i we can obtain them by solving the following equation:

$$max \quad W(\alpha) = \sum_i \alpha_i - 1/2 \sum_i \alpha_i y_i z_i$$

$$s.t. \quad 0 \leq \alpha_i \leq C \quad \wedge \quad \sum_i \alpha_i y_i = 0 \quad \forall i$$

x_i is the input y_i is the corresponding target value and C is the penalty factor; if C is too large we risk overfitting, which causes increased costs for non-separable points. Though the generation of many support vectors reduces the training error, it also makes the model overly complex. The optimal kernel parameters can be calculated using the Fisher discriminant method to find linear combinations of features. These are specific technical details related to our classifications used in this chapter.

As mentioned earlier, in our modelling we used the SMO version of SVM implemented by Hall et al. (2009). The algorithm was explained in detail by Platt (1998), Keerthi et al. (2001) and Hastie and Tibshirani (1998); Platt (1998) also provided the implementation details. Basically, the procedure eliminates one of the Lagrange multipliers, and transforms the optimization model above to a quadratic minimization model with one variable, while discarding the required direct knowledge of a threshold in each iteration.

4.1.4 LSA and LSI

Humans are able to understand the topics of a complex text, and we think that is somewhat related with the intuition behind co-occurrences of the terms. The idea behind automatic semantic analysis is then to simulate and automate these crude intuitions. For example, when mæy or "wine" co-occur with dɔrânæ-ʃabâb or "young days" there is a different semantic relation than when they co-occurred with kəʃam raxt bə məyχânə or "let me take my being to the bar". In the former, the theme is more materialistic and joyful, whereas the latter sentence references a broader philosophical way of life.

Deerwester et al. (1990) developed Latent Semantic Indexing, and used the concept of Singular Value Decomposition (SVD) of document-term matrices. SVD is central to LSI

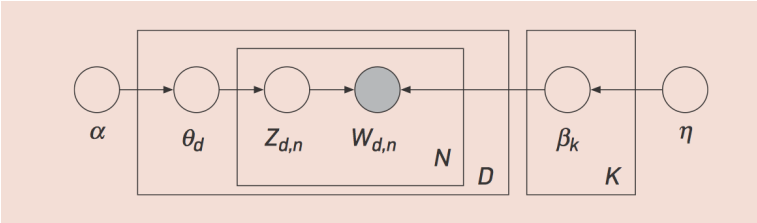


Figure 8: LDA graphical model

and is based on eigenvalues and eigenvectors. With any n x n non-zero matrix A, that is multiplied by special scalars (eigenvalues), n-dimensional eigenvectors stretch, contract and reverse, though they do not change direction.
Therefore, if we rewrite any square diagonalizable matrix A as the multiplication of its eigenvectors, matrix S, and its transposition S^T on either side of the diagonal matrix of its eigenvalues, we can transform and normalize the diagonal matrix Λ:

$$\boxed{A = S\Lambda S^T}$$

Since it is already composed of perpendicular vectors of length 1 (orthonormal vectors), we refer to this transformation as singular values and call this decomposition of the matrix A *singular value decomposition*. We capture latent dimension in the new matrix where we group synonyms together, but as the semantic grouping is not explicit they are latent concepts. We use this technique to determine the semantic term relations within a document, which allows us to rank documents with closer semantic relations higher, given the specific terms we captured have the significant eigenvalues or *singular values*. See (Noorinaeini and Lehto, 2006) for a more complete treatment of the SVD procedure.

4.1.5 LDA

Figure 8 from (Blei et al., 2003) depicts a graphical model of LDA[27]—we shaded the observed nodes and hid the rest. Plate N indicates the replication of words within documents. Plate D is the collection of documents within the corpus. θ is a matrix of mixture distribution of D documents over K topics from a Dirichlet prior parameterized by α and β is the distribution of K topics over the terms W from a Dirichlet prior with parameters η. θ and β are to be estimated. Automatically extracting the main topics of documents is known as topic modelling (Blei, 2012). We use LDA, a probabilistic model, and its associated algorithm to identify co-occurring terms within a document and their

[27]Latent Dirichlet Allocation

posterior probability. Probabilistic LSI [28] is, in essence, the parent of LDA, and Blei developed it to improve LSI. According to Blei (2012), LDA is very similar to PCA from the matrix factorization point of view. PCA reduces the dimensionality without losing much of the information by keeping important data. The topic labels are latent and internal to the LDA model, therefore the topic terms belong to and are grouped under their latent topic.

The LDA technique is part of the work to extract the underlying theme by quantifying term-topic-document relations. It creates topics based on the distribution of words in the corpus, then weighs the documents' relevance to their latent topics.

LDA is more difficult to stipulate, perhaps due to the combination of joint conditional probabilities of hidden variables and observed variables. The well-known mathematical relations by (Blei, 2012) shown in Figure 8. Here is Blei's equation:

$$P(\beta_{1:k}, \theta_{1:D}, Z_{1:D} \mid W_{1:D}) = P(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D})/P(W_{1:D})$$

where topics have a $\beta_{1:k}$ distribution over the terms; $\theta_{d,k}$ is the topic proportion of document d with respect to the topic k. $Z_{d,n}$ is the topic assignment of term n with respect to document d. $W_{d,n}$ is the observed nth term in document d.

We call this equation *posterior probability*, and its denominator marginal probability or evidence, which is exponentially large due to topic structures. The probability of a term belonging to a document is proportionally much larger than its probability belonging to a specific topic. Hence, we approximate the posterior probability with *Gibbs* sampling rather than variational algorithms. *Gibbs* method is a sampling technique, which gathers approximated data, based on a multivariate probability distribution and a Markov chain Monte Carlo algorithm (Carlo, 2004). Sampling algorithms builds a sequence of random variables, each conditional on the previous one (Griffiths and Steyvers, 2004). In variational methods, we replace the sampling inference with parameterized distribution of hidden structures, to find the closest point to the posterior probability. This is an optimization problem, and is an active aspect of the research to investigate whether variational or sampling methods are better suited for the topic modelling task. The LDA used in (Ghayoomi and Momtazi, 2014) favours the *Gibbs* sampling.

For LDA implementation, we have used the ldavb.py script by M. Hoffman [29] and the *Gensim*[30] library that utilizes *Gibbs* sampling from MALLET [31] in their LDA implemen-

[28] Latent Semantic Indexing

[29] http://www.cs.princeton.edu/~mdhoffma

[30] https://radimrehurek.com/gensim/models/ldamallet.html

[31] http://mallet.cs.umass.edu/

tation (Řehůřek and Sojka, 2010). The latter implementation with Gibbs sampling has proven the best for our purposes [32].

4.1.6 LDA-based topic probabilities

The objective is to generate topics and calculate the probability of the topic terms in a ghazal. To generate LDA features for the entire corpus of poems, we performed the following steps: Munková et al. (2013) studied the influence of stop-words on the quality of text processing and concluded that stop-word removal did not influence the textual pattern discoveries. Al-Shargabi et al. (2011) concluded that stop-words removal improved SVM's accuracy when classifying Arabic text. We have removed stop-words for most of our classification experiments but have kept them for visualizations. We are not claiming of any particular classification impact attributed to stop-words or the lack thereof.

1. Filter out the stop words; (They tend to not help in the classification, since they occur in all the classes.)

2. Remove words that occur only once; (They tend to not help in the classification, because they will appear only in training or in the test data.)

3. Create a dictionary structure (a hash table) of all words in the corpus, to use for the initialization of the LDA model; The dictionary was used to map the IDs to words in the vocabulary.

4. Create the bag-of-words (BOW) matrix;

5. Create the TF-IDF model using the BOW matrix;

6. Train and initialize the LDA model using both the dictionary and the associated TF-IDF values;

The time complexity needed to compute the features and to train the classifier is low. Once the LDA model is prepared, we can use it to produce topic probabilities for every ghazal. The result is a set of probability values with the same number of topics as we chose for our LDA model. We mostly used very few topics such as 5, 6 or 10 to support human stipulation and intuition when reviewing the topics.

[32] https://radimrehurek.com/gensim/models/ldamodel.html

4.1.7 Similarity Features

Houman's conceptual guidelines about the classification of Hafez's ghazals made us realize that our knowledge representation should capture and represent the meaning and concepts that Hafez buried in the poems. Therefore, ML methods should be concerned with semantics, and be capable of capturing as much meaning as possible. In search of semantic ML methods, we also come across LSA (Landauer et al., 1997). LSA is an unsupervised ML method that can provide the means to make decisions about the similarity of words in their contexts.

The idea behind LSA raises the question of how to mimic the human understanding of texts based on the association of the words. As mentioned, we intend to get close to human ability to determine the subject of a text and conceptually relate LSA representation to a classification algorithm.

LSA also allows us to find the instance similarities via the Cosine similarity measure used in vector space modelling. We created our Cosine similarities using either LSI or LDA models in conjunction with the previous dictionary. We indexed for the entire corpus, then used the index to calculate the similarities to any unlabelled ghazal. We based our LSA model on the whole corpus and created the BOW, dictionary and TF-IDF to train or initialize the model. Then, using the index created for the Cosine matrix, we calculated the similarity values for every unlabelled ghazal by iterating through the index. This allowed us to determine an unlabelled ghazal's similarity values to all the training ghazals in the index. The procedure can be summarized as follows, which is similar to those in Section 4.1.6 with additional steps:

1. Filter out the stop words.
2. Remove words that occur only once.
3. Create a dictionary structure of all words to use to initialize the LDA model.
4. Create the bag-of-words matrix.
5. Create the TF-IDF model using the BOW matrix.
6. Train and initialize the LSI or LDA model using both the dictionary and the associated TF-IDF.
7. Create the Cosine Similarity Matrix and the index using the LSI or LDA model.

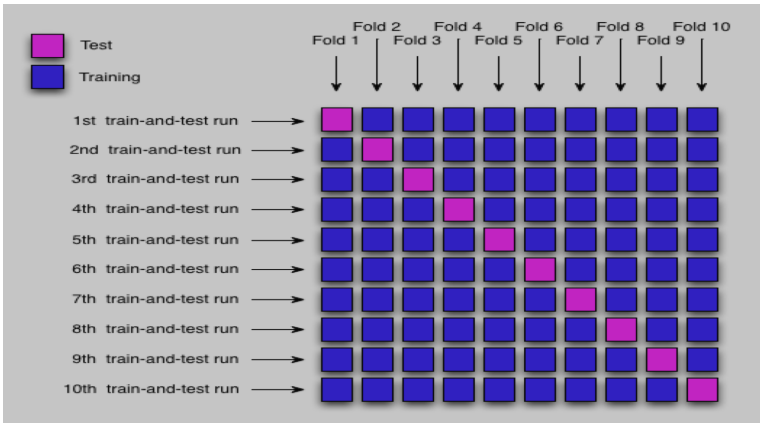


Figure 9: Ten-Fold cross-validation

8. For every unlabelled ghazal, iterate through the index and calculate the similarity values for each ghazal in the corpus.

4.1.8 Evaluation method

We employed the stratified 10-fold cross-validation method (Baccianella et al., 2010). Stratified cross-validation means that each fold contains the same proportions of class labels, and the evaluation method partitions the data set into 10 portions. We created the ten sets of evaluations for each experiment, each of which uses nine subsets for training and one to test the data and summarize the ten evaluation results. This allowed us to calculate *accuracy*, *precision*, *recall* and *F measure* using this *stratified 10-fold cross-validation* method for each class. Figure 9 from (Baccianella et al., 2010) depicts this idea well.

4.1.9 Visualization method

The objective of the visualization is to better reveal some of the nuances mentioned above such as the topic terms of each class. This way, we can depict the interactions of the main topics of the ghazal on its defined or predicted class. The overall shapes of the resulting networks appeared to have tangible and comparable indications, which provided us with some insights regarding the topic terms of the LDA models. We used Principal Component Analysis (PCA) to derive presentable graphs of the interrelations of LDA topics. Linear algebra discusses the eigenvectors and eigenvalues and probability discusses the covariance, and these form the basis of the PCA, which uses orthogonal or perpendicular relations of the vectors to help surface the patterns of data, particularly

in high dimensions. In fact, PCA is a nonparametric technique to reduce the dimensions without excessive loss and keep only the relevant information. PCA is conceptually similar to SVD.

To create the PCA model, we first subtract the mean of each dimension from the dimension components and calculate the covariance matrix. We then calculate the unit eigenvectors and eigenvalues of the covariance matrix, which maintains the main attributes and characteristics of the data. The principal component is the eigenvector with the highest eigenvalue. The dimension reduction mechanism discards the dimensions with low eigenvalues, scores them accordingly and keeps the vectors we consider important.

These key components shape the feature values that we multiply by the mean-adjusted data, and the approach provides almost the same as the original data regarding our chosen vectors. This framework gives us a perfect platform for our graphical representation of the data transformed into two-dimensional Euclidean line distances, which inherently function as proxies to expose the differences and similarities of patterns among the data sets (for more detail, refer to (Jolliffe, 2002)). We use the earlier steps of initiating the LDA model by the TF-IDF of BOW, before passing the vectored topics to the PCA object to create graphs.

The first set of graphs depicts LDA-PCA topics by their ranges as 2D shapes, and the second set shows how we use LDA-PCA to create a scatter plot of the topic ranges, though we used the top terms as points. The third set of graphs shows LDA-PCA objects as a network of topics, with the edges weighted by the inverse of node correlations. The values on the x and y axis are all the 2D PCA dimensions we reduced from their LDA vector space. We can quickly and independently create and train an LDA model by performing the first five steps. Although the following technical process steps for visualization are decoupled from those for modelling, the first 6 are almost identical between the two[33]. Refer to the steps in section 4.1.6 to compare. The time complexity of the algorithm is low.

1. Filter out the stop words.
2. Remove words that occur only once.
3. Create a dictionary structure of all words input from the corpus, to be used for initialization of the LDA model.

[33]https://gist.github.com/tokestermw/3588e6fbbb2f03f89798

4. Create the bag-of-words matrix and collect document frequencies.

5. Create the TF-IDF model using the BOW matrix and calculate the IDF weights.

6. Train and initialize the LDA model using both the dictionary and the associated TF-IDF to evaluate the perplexity, and perform 50 iterations with a convergence threshold of 0.001000, as per the default setting of the gensim library.

7. Use a dictionary vectorizer to transform the LDA topics to vector space. Dictionary or hash table maps words of the documents for the transformation function to set up the vector elements properly.

8. Create a PCA model using the normalized transformation of the topic vector to the array.

9. Create a squared matrix form of the PCA object by using a Euclidean metric as distance.

10. Create a graph object using the PCA squared form attributes as edges while applying a constant weight factor.

11. Draw the graph as a network of nodes, edges and labels. Refer to results and visualization 6.1.3

4.1.10 Classification Experiments

In this section, we explain the seven experiments we conducted, each of which consist of many sub-experiments. The main experiments are two-stage models: We created the training features by unsupervised methods in the first stage, then used the labelled vectors to train the SVM classifier in the second stage. We present the state-of-the-art Hafez classifier outcome we developed using *LDA-Cosine similarity* as features for all six Houman classes. We initially used the BOW as features, then later we applied LDA-topic-term-probability factors from both the Persian and the English sections of the corpus as features.

Before any classification attempt it is important to define a baseline, and the most fundamental baseline measure is the proportion of the largest class size to all the data; this is referred to as ZeroR classifier in Weka (Hall et al., 2009). This classifier puts everything in the majority class, so to be useful our classifiers must have better performance than baseline. Our champion classifier outperforms the majority class baseline (31.7%) and

the random guess baseline (58.2%).
Our Hafez training corpus, based on Houman's classification, includes the instance counts. To calculate the baseline, we observed that the 'Maturity' class had the highest number of instances (79), and divided this by the total number of training instances (249). Therefore, the accuracy of the baseline classifier for six classes is the percentage of correctly classified instances; that is 79/249 = 31.7269%. The baseline classifier for three classes has the accuracy 145/249 = 58.2%. As mentioned, we used the 10-fold stratified cross-validation testing mode.

4.1.11 Baseline and Bag-of-Words evaluation

The first feature set we applied to train the SVM classifier was the BOW, with 2,083 instances with 5,411 attributes. This improved the baseline with the following results:
Correctly classified instances: 93 (**37.3494%**)
Mean absolute error 0.2544
Root mean squared error 0.3584
Overly high regression residuals may be attributed to the non-normality of the errors, therefore we weigh in the accuracy in a discrete classification.
These findings indicate a 5.6% increase over the baseline accuracy. Due to the higher volume in Maturity (class *c*), most other classes tended to fall into it. We labelled the classes *a*,*b*,*c*,*d*,*e* and *f*, in chronological order [34]. As shown in the confusion matrix, only 33% of class d is correctly classified, and the rest is under class c. Classes b, e and f are all classified incorrectly, mostly as c or d. Of most concern is the imbalance of the

a	**b**	**c**	**d**	**e**	**f**	**<- classified as**
2	0	32	3	1	0	a
1	0	17	6	1	0	b
0	0	69	10	0	0	c
0	0	44	22	0	0	d
0	0	21	7	0	0	e
1	1	10	1	0	0	f

Table 3: Confusion Matrix: Persian BOW

[34] Youth=a, PostYouth=b, Maturity=c, MidAge=d, Before Elderly=e, Elderly=f

classification, as most instances are classified in c and d. The precision measures for each class are shown in Table 4. The interchangeability of classes c and d indicates that combining them might be a good idea in further experiments. In the next experiment, we

TP Rate	FP Rate	Precision	Recall	F-Measure	AUC Area	Class
0.053	0.009	0.5	0.053	0.095	0.644	a
0	0.004	0	0	0	0.443	b
0.873	0.729	0.358	0.873	0.507	0.571	c
0.333	0.148	0.449	0.333	0.383	0.651	d
0	0.009	0	0	0	0.625	e
0	0	0	0	0	0.485	f

Table 4: Performance Matrix: Persian BOW

added some English translations to our training corpus and repeated the BOW features in both languages together to train the SVM classifier. The correct classification became 39.759%, which is a 2.4% improvement.[35] As we see in the confusion matrix 5, class a had the least improvement. However, the overall improvement leads to better recognizing between class c and d, while classes e and f are not recognized.
Correctly classified Instances: 99 (**39.759%**)
Mean absolute error 0.255
Root mean squared error 0.3592

As mentioned, the distribution of evaluation results showed that the bulk of instances classified by the SVM model fell into classes c and d. Therefore, in the next experiment we decided to combine adjacent classes and train the classifiers with three classes: a and b, c and d and e and f, resulting in new classes a', b' and c'.
The BOW result for this approach is **60.6426%** and out of 249 instances 151 were correctly classified .
Root mean squared error: 0.4471
The detailed weighted average measures across all three classes are: TP Rate=0.606,

[35]Statistical significance tests would need to be done to see if the improvement is significant, for this classifier and for the other results reported in this thesis. We did not consider necessary to run statistical significance test for the small improvements from one experiment to another. The final classifiers we propose achieved reasonable accuracy, with large enough improvements over the initial ones.

a	b	c	d	e	f	<- classified as
3	0	25	9	1	0	a
1	0	13	10	1	0	b
1	0	62	16	0	0	c
0	0	32	34	0	0	d
2	0	15	11	0	0	e
0	0	10	3	0	0	f

Table 5: Confusion Matrix: English and Persian BOW

FP Rate=0.527, Precision=0.596, Recall=0.606, F-Measure=0.493 and AUC Area=0.54. However, the distribution of the classified instances by the evaluation is still (predictably) bulked in the center.

a'	b'	c'	<- classified as
7	55	1	a'
2	143	0	b'
2	38	1	c'

Table 6: Confusion Matrix: 3 classes Persian BOW

4.1.12 Semantic Features

At this point, we have two challenges: improving classification accuracy, and upgrading the class distribution balance so classes a, e and f are better recognized by the classifier. We employed semantic features to add more diverse meaningful features. In this case, we empirically determined that initiating the LSI module with six topics gives the optimal effect of ranges from 3 to 20 topics. These are the numbers of topics chosen for the LDA-driven features. The iterations with larger number of topics did not necessarily improve the classification. After initializing the LSI module with the training data of three classes, the model calculates the probability measures for each of the six LSI internal topics for the 2-stage method of transformed TF-IDF of BOW for each ghazal. See transformation details in Section 4.1.2. We then combined these with the original BOW features, before training the SVM classifier.

Compared to the pure BOW, the final evaluation of 10-fold cross validation of this method

achieved a 6.02% improvement of correctly classified ghazal instances to the right Hafez periods:

Correctly classified instances: 164 (**66.6667%**)

Mean absolute error: 0.3244

Root mean squared error: 0.4198

Examining the confusion matrix below, we see that the classifier is slightly better at recognizing class a' than the SVM classifier trained with only pure BOW features. However, though the final classifier SVM model is improved, our classifier is still weak at recognizing class c'. So, we need to find a way to tackle the imbalance problem.

a'	b'	c'	**<- classified as**
20	42	1	a'
2	142	1	b'
3	37	1	c'

Table 7: Confusion Matrix: 3 classes Persian BOW + LSI distributions

The accuracy matrix shows that the highest F-measure is again concentrated in class b'. Nevertheless, the class a' F-measure is improved and has the highest precision (90%). This may be due to the fact that common terms exist in all classes and these semantic features are weak to discriminate when the same terms have deeper meaning in the later classes.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AUC Area**	**Class**
0.349	0.013	0.898	0.349	0.503	0.745	a'
0.99	0.764	0.643	0.99	0.78	0.613	b'
0.012	0.005	0.333	0.012	0.024	0.553	c'

Table 8: Performance Matrix: Persian BOW + LSI distributions

As is evident, the precision for class c' is the lowest, and the F-measure is also extremely low.

4.1.13 Latent Dirichlet Allocation Similarity Measure

After many trial and error experiments, we designed a technique that improved the classification distribution balance. More specifically, we developed features that can be used to train the model to better distinguish class c'; its instances are reference vectors to calculated similarities for all instances.
This method not only creates better classification balance, but also helps improve the performance by up to 2.7%.
Correctly classified instances: 171 (**69.3878%**)
Mean absolute error: 0.3158
Root mean squared error: 0.4095
In the confusion matrix we found that 13 of 41 instances of class c' were correctly recognized.

a'	b'	c'	**<- classified as**
23	39	1	a'
1	143	1	b'
1	27	13	c'

Table 9: Confusion Matrix: 3 classes Persian BOW + LSI + LDA similarity

The detailed performance shows that for class c' we achieved a 98% precision and a 50% F-measure. This low recall may be attributed to the imbalance and gravitation towards the center class, and as we see the largest class is mostly classified correctly, causing the number of false positives to be much smaller than false negatives.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AUC Area**	**Class**
0.365	0.015	0.885	0.365	0.517	0.759	a'
0.99	0.635	0.645	0.99	0.781	0.678	b'
0.333	0.002	0.976	0.333	0.497	0.706	c'

Table 10: Performance Matrix: Persian BOW + LSI + LDA similarity

4.1.14 Classification of the Bilingual Corpus

In this experiment, we combined all the techniques we had learnt so far and achieved the best possible outcome; that is, a highly accurate Hafez classifier. We also used the bilingual corpus with the best translations we had found, and combined them with the Persian corpus we employed in previous experiments.
We created all the features using this corpus and carefully combined them while maintaining the ghazal order and index. First, we examined the SVM results and performance using BOW.
Correctly classified instances: 162 (**65.0602 %**)
Mean absolute error: 0.3356
Root mean squared error: 0.4329

a'	b'	c'	**<- classified as**
29	34	0	a'
12	133	0	b'
14	27	0	c'

Table 11: Confusion Matrix: 3 classes Persian/English BOW

The detailed performance shows that with class b' we achieved 92% recall.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AUC Area**	**Class**
0.46	0.14	0.527	0.46	0.492	0.625	a'
0.917	0.587	0.686	0.917	0.785	0.639	b'
0	0	0	0	0	0.486	c'

Table 12: Performance Matrix: Persian/English BOW.

Next, we kept the primary BOW features and created the LDA based values for each ghazal. We find the results in the confusion matrix in Table 13 and we used these two sets of BOW, LDA features to train the SVM model.
Correctly classified instances: 183 (**73.494%**)

Mean absolute error: 0.2811
Root mean squared error: 0.3647

a'	b'	c'	**<- classified as**
0	63	0	a'
3	142	0	b'
0	0	41	c'

Table 13: Confusion Matrix: 3 classes Bilingual BOW + LDA distribution factors.

The detailed performance shows that for class b' we have been able to achieve a 98% recall.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
0	0.016	0	0	0	0.538	a'
0.979	0.606	0.693	0.979	0.811	0.563	b'
1	0	1	1	1	1	c'

Table 14: Performance Matrix: Persian/English BOW + LDA distribution values

In Table 13 all class a' instances were classified as class b', so we next calculated the similarity features for all, keeping the class a' instances as vector references. As a result, we observed the performance of SVM when its training data has the bilingual BOW, LDA factors, and the new similarity values. We found that this combination increased the accuracy up to 86%. The confusion matrix is in Table 15.
Correctly classified instances: 215 (**86.354%**)
Mean absolute error: 0.2267
Root mean squared error: 0.2802

The detailed performance shows that for class a' we achieved 95% recall see Table 16.

a'	b'	c'	<- classified as
60	2	1	a'
13	125	7	b'
5	6	30	c'

Table 15: Confusion Matrix: 3 classes Bilingual BOW + LDA + Similarity values

TP Rate	FP Rate	Precision	Recall	F-Measure	AUC Area	Class
0.952	0.096	0.769	0.952	0.851	0.788	a'
0.862	0.077	0.940	0.862	0.899	0.861	b'
0.732	0.038	0.789	0.732	0.759	0.787	c'

Table 16: Performance Matrix: Persian/English BOW + LDA + Similarity values

4.1.15 A more Fine-Grained Classification

In this experiment, we continued using the English/Persian corpus to train our SVM model, but we only included Houman's classes 3 and 4, in order to better predict to a more granular level, meaning to break the class b' into c and d. Considering the best model results in the previous experiment, we predicted that all unclassified ghazals by Houman belong to the Maturity and Middle-age classes, as shown here in c and d. In this experiment, we intended to further distinguish the two, and thus prepared a training corpus comprised of only the two granular classes c and d: our first attempt is to see the performance of the SVM with BOW features.

The baseline is 54.4% (= 79/145), and our modelling improves on the baseline by almost 14%; that is, SVM has 68% accuracy. However, there are important technical differences with the previous experiments, since in all of them, the last one was the best performer. However, in experiment groups 2 and 3, the training set was included the BOW, LSI and LDA values. In the case of experiment group 4, we added the LDA-Similarity values.

In other words, we always kept the BOW as part of the feature space, which allowed us to achieve the best performance in that group. Notably, in this experiment, unlike the previous attempts, we were able to achieve top performance without the BOW directly participating in the training feature space. This last group of experiments indicates that, given the conditions of our Hafez task, LDA-Similarity values are very powerful predictors. The BOW results are:

Correctly classified instances: 99 (**68.2759%**)
Mean absolute error 0.3172
Root mean squared error 0.5632
The confusion matrix and accuracy measures are as follows, Tables 17 and 18:
The detailed performance shows that for both classes we achieved a 79% and 56% recall,

c	*d*	**<- classified as**
62	17	*c*
29	37	*d*

Table 17: Confusion Matrix: 2 classes Persian/English BOW

for c and d respectively.

TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**AUC Area**	**Class**
0.785	0.439	0.681	0.785	0.729	0.549	*c*
0.561	0.215	0.685	0.561	0.617	0.664	*d*

Table 18: Performance Matrix: Persian/English BOW

LDA-values have always provided powerful performance upgrades, but only when directly combined with the BOW in the training set for SVM. As shown below, with the LDA-values as stand-alone training sets for SVM, our classifier performs marginally better than the baseline. However, compared with BOW the performance degrades significantly, by about 13%.
Correctly classified Instances: 81 (**55.8621%**)
Mean absolute error: 0.4414
Root mean squared error: 0.6644

c	*d*	**<- classified as**
49	30	*c*
34	32	*d*

Table 19: Confusion Matrix: 2 classes Bilingual LDA distribution values

The detailed performance shows average recall for both classes when BOW features are not used.

The LDA distribution values used to train SVM have weaker classification performance

TP Rate	FP Rate	Precision	Recall	F-Measure	AUC Area	Class
0.62	0.515	0.59	0.62	0.605	0.544	*c*
0.485	0.38	0.516	0.485	0.5	0.522	*d*

Table 20: Performance Matrix: Persian/English LDA distribution values

than with BOW alone. However, when the similarity values are added to the LDA values as training data, we achieve the best possible performance using the SVM model. Another difference with this experiment is we added the LDA similarity values as vector references for ghazals in the c and d classes. In other words, we trained or initialized the LDA-similarity model for all participating Maturity and Mid-Age classes in training, then calculated the LDA similarity values for these training ghazals. This way, we ensured there was no biased information for any specific class to participate in the training features. In the experiment below, combining the LDA with the LDA-similarity values achieved the highest performance:

Correctly classified instances: 134 (**92.413%**)

Mean absolute error: 0.3358

Root mean squared error: 0.3809

c	*d*	**<- classified as**
75	4	*c*
7	59	*d*

Table 21: Confusion Matrix: 2 classes Bilingual LDA distribution + Similarity values

The detailed performance shows that for both classes we have been able to achieve above 90% F-measure see Table 22.

The predictions of this two-phased model will be used for the analysis in Chapter 6, and we will compare them with the predictions of our champion model discussed below.

TP Rate	FP Rate	Precision	Recall	F-Measure	AUC Area	Class
0.949	0.106	0.914	0.949	0.932	0.893	*c*
0.893	0.051	0.936	0.893	0.915	0.940	*d*

Table 22: Performance Matrix: Persian/English LDA distribution + Similarity values

4.1.16 LSI Similarity vs. LDA Similarity Features

In this experiment, we take advantage of similarity values, and in this subsection, we get a better sense of the features category by isolating them and comparing the results. We conducted two main experiments with the Cosine similarity features in isolation. Refer to the only results for LDA features that are shown here: Table 23.
In both experiments, we used the bilingual corpus labelled with the set of three classes and the chronological grouping of Houman's original six classes, and derived the Similarity Matrix values differently. In the first experiment, we calculated the similarity weights based on an LSI model for all ghazal instances with respect to the sample ghazals coming from class two the largest class of the corpus. In the second experiment, we performed the same procedure, except we calculated the weights using the LDA model, which surpassed the accuracy of classification using LSI features.

4.1.17 Classification of all classes

In this experiment, we classified the entire corpus with Houman's original six classes, as opposed to incorporating the chronologically adjacent pairs into one class to reduce the classes to three. In the second group of experiments above, the intention was to reduce the number of classes to improve on the SVM classification. In experiment group one, we noticed a significant 20 point performance improvement when using BOW features and the three classes; it went up approximately from 40% to 60%. With the similarity feature values performing well, we returned to the original corpus tagged with six classes of Hafez ghazals and applied the top performing method we just used. To avoid potential biases, we calculated the similarity-feature values with respect to all ghazals for every ghazal in the training and the test data.
Correctly classified instances: 197 (**79.1164%**)
Mean absolute error: 0.2227
Root mean squared error: 0.311

a	b	c	d	e	f	<- classified as
35	0	2	1	0	0	a
3	16	4	2	0	0	b
1	0	65	11	2	0	c
0	3	8	52	3	0	d
0	1	2	6	19	0	e
0	0	0	1	2	10	f

Table 23: Confusion Matrix: Persian LDA-Similarity

The following is the detailed performance matrix.

TP Rate	FP Rate	Precision	Recall	F-Measure	AUC Area	Class
0.921	0.019	0.897	0.921	0.909	0.869	a
0.64	0.018	0.800	0.64	0.711	0.853	b
0.822	0.094	0.802	0.822	0.813	0.795	c
0.787	0.0115	0.712	0.787	0.748	0.715	d
0.678	0.032	0.730	0.678	0.704	0.692	e
0.769	0	1	0.769	0.869	1	f

Table 24: Accuracy Matrix: Persian LDA-Similarity for 6 classes

There was no need for English in the corpus for this experiment. This is the champion classifier that we used to do our best predictions on the unlabelled ghazals.

4.1.18 Summary Highlights

We observed that the following features or methods increased the accuracy of our classifiers: The LSI or LDA-driven standalone similarity features should provide strong enough training features. Therefore, we created the training data with only normalized similarity features, once with LSI and once with LDA for classes a',b',c'. The former resulted in 62% accuracy, while the latter reached 86%; to our surprise, this similarity feature alone demonstrated a very powerful training data for SVM. We reviewed our

program design and methodology multiple times and verified that there was no overfitting or bias, by reviewing measures such as recall and precision. Since we had found this powerful method, in the final set of experiments we used the set of six classes and prepared the training data based only on the LDA-driven Cosine similarity features on all classes, and achieved an accuracy of almost 79%. We then applied the final model predictions for to visualization and analysis of the results and had some of the unlabeled ghazals validated by the two experts.

1. Reducing the number of classes improved the performance.
2. Including the English with the Persian corpus always improved the performance.
3. LSI or LDA semantic features in conjunction with the BOW improved the performance; LDA distribution features were more powerful than those of LSI.
4. LDA Cosine similarity feature values performed better than LSI.
5. Our LDA-Cosine similarity features produced by balancing the confusion matrix were the most powerful.
6. The more anchors used to calculate similarity features achieved the best classification performance, while similarity references (larger dimensions in the feature set) improved the accuracy.

4.2 Measuring Inter-annotator agreement (Kappa) and Coherence

Cohen's Kappa (Fleiss et al., 1969) measures the agreement between two sets of labels, generated by classification or clustering. Hubert (1978) later refined the index and used the *weighted* Kappa as a bilinear permutation assuming marginal frequencies of responses were fixed. p_o and p_e are probabilities of the relative observed and hypothetical agreements among raters respectively:

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

In our case of comparing Raad and Houman annotations, Spearman correlation=0.897, (p value=6.18e-178) and *kappa* with linear weights was 0.68, which is substantial (McHugh, 2012). In linear weights vs. quadratic, the distance between classes and the number of categories are not squared.

Class	Coh_{umass}	Coh_{uci}	LogPerplexity	
3 Cls Avg	-10.37	**0.52**	-10.12	
6 Cls Avg	-7.61	**0.55**	-8.67	

Table 25: Houman Labels of Three and Six Classes: Coherence

We used several coherence measures proposed by (Röder et al., 2015), which were based on *pointwise mutual information* and confirmation measures. (Mimno et al., 2011) had used $log(PMI)$ in the definition of *coherence*, which drew from the smoothed conditional probability of asymmetric confirmation measure of top words per topic. (Řehůřek and Sojka, 2010) implemented them in *gensim* Python library. Inspired by PMI, Newman et al. (2010) developed the UCI measure and Mimno et al. (2011) developed the UMass measure of coherence. $P(w)$ is the probability of word tokens. The subscripts mean University of California at Irvine and University of Massachusetts respectively. The coherence closer to 1 would be the better.

$$Coh_{UCI} = \frac{2}{N(N-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{N} \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \tag{3}$$

$$Coh_{UMass} = \frac{2}{N(N-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \tag{4}$$

The Houman labels with 3 and 6 classes have higher coherence, shown in Table 25, than when we merge them into 4 classes, to compare with Raad's [36] (Raad, 2019) classification [37], in Table 26. The higher the Log Perplexity is, the better when comparing LDA models (Hoffman et al., 2010). To have a homogeneous comparison between the two independent scholars' annotations, we merged the Houman 6 period chronological classifications into 4 (class 1 includes classes a and b, class 2=c, class 3=d,class 4 includes classes e and f). The LDA-driven coherence (Coh_{cv}) of Houman (0.51) is still higher than that of Raad (0.49) in Table 26.

4.2.1 Classification Refinements

The objective of this section is to measure the impact of inconsistent or disagreed upon instances on the change in coherence for each class, which can guide labelling to refine the performance in classifications.

[36] Contemporary Hafez scholar.

[37] We used LDA-driven Coherence and Log Perplexity using gensim Python library.

		Raad			Houman	
Class	Coh_{umass}	Coh_{uci}	LogPerplexity	Coh_{umass}	Coh_{uci}	LogPerplexity
Avg	**-8.87**	**0.49**	**-7.72**	**-8.56**	**0.51**	**-9.18**

Table 26: Raad and Houman Labels of Four Classes: Coherence

4.2.2 Preprocessing

In our preprocessing, we removed the stop-words and the tokens that occurred only once, refer to Section 4.1.6. 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 then built the TF-IDF vectors accordingly. We initialized LSI, LDA[38], Log-Entropy (Lee et al., 2005) and Doc2Vec (Le and Mikolov, 2014) objects using the Persian section of our corpus as training. Doc2Vec is based on Word2Vec which are word embeddings or distributed representations of words using NN models (Mikolov et al., 2013); both are extended to build Continuous Bag-of-Words (CBOW) and the Skip-Gram algorithms. We used the gensim library (Řehůřek and Sojka, 2010) and used HAZM[39] Python library for Persian preprocessing tasks, such as *tokenization*, *normalization* and *lemmatization*.

4.2.3 Labelling Inconsistency Management

Inter-annotator agreement has been used to refine classifications. For example, Wiebe et al. (1999) used kappa results to inform annotations in an iterative manner and were able to both improve the classifier's accuracies and inter-agreements among the annotators. If multiple classifiers predicted, classified, labelled, or voted differently for an instance record or for a unit of training material, then the question becomes how we can decide its class better. If we wanted to re-annotate such instance as training material, naturally we consider the effects of such items on the quality of training material for machine learning and predictions. We also used coherence as an index of within-class relatedness. Hence we considered the semantic effects of either changing the label of an instance poem or excluding it from the training material when conclusive or otherwise

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

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

disregarded to enforce any changes in the label if the improvement in coherence was very small. In this example, we showed that the isolated inconsistencies between two independent labelling systems of Houman and Raad could provide us with clues to improve and produce a third but refined training material for classification. A similar concept is used in a bottom-up clustering in which maximizing the mutual information between adjacent classes made up the algorithm (Manning et al., 2008); during the iterations, they only merged the two clusters that produced a minimum loss in mutual information.

$$(c_{n1}, c_{n2}) = \underset{c_i, c_j}{\operatorname{argmin}} MI\text{-}loss(c_i, c_j) \tag{5}$$

We extended the notion of loss in mutual information in Equation 5 and used LDA-driven coherence change per class to measure the impact of inconsistent labels. We calculated the difference in coherence among the union of the inconsistent poems with their corresponding two classes that our scholars had labelled differently. We then measured and anticipated the semantic impact of the inconsistent segment accordingly, by numerical analysis of the rate of change in coherence among the segment's memberships. See Equation 6. The interplay of coherence measurements guided and informed the training-data decisions and their make-up. For example, when we added the inconsistent segment to either class voted by classifiers, and it deteriorated the coherence in both cases, then using *DeltaSem* index defined in Equation 6 we compared the rate of change in both cases to measure the difference. In other words, the index informed the direction of change. If both ΔCoh were positive, it meant that the inconsistent segment impaired either class. Further, if the *DeltaSem* was small ($\delta \to 0$), it meant that the inconsistent segment almost equally impaired both classes; therefore it was better to exclude the inconsistent segment from adjacent training classes altogether.

$$DeltaSem : |\Delta Coh(c_{incons.}, c_{youth}) - \Delta Coh(c_{incons.}, c_{Mid-age})| \overset{?}{\to} \delta \tag{6}$$

For example, to demonstrate that this procedure of identification of culprit poems can refine the quality of the training data and hence improve classifications. We took the items out of the multiple categories of inconsistencies for which Houman and Raad had labelled *Youth* and *Mid-age* respectively, and compared their change in coherence. As we see in Table 27, the exclusion of inconsistent instances improved coherence for both classes Youth and Mid-age by an almost equal amount of 3 percentage points. Also, the difference in coherence change was small, consistently confirmed that the odds of improvement in keeping inconsistent instances in either class were small: $\delta \to 0.008$.

Class	Coh_{uci1}	Coh_{uci2}	δ
Houman Cls1	0.355	0.388	0.033
Raad Cls3	0.395	0.420	0.025

Table 27: Raad and Houman Labels Consistency Improvements: Coherence

Feature	SVM:Acc.,F1	Reg:Acc.,F1	DT:Acc.,F1	NN:Acc.,F1	RF:Acc.,F1
WE	0.48, 0.31	0.28, 0.27	0.4, 0.41	0.48, 0.31	0.40, 0.35
WE_{DM}	0.32, 0.29	0.28, 0.27	0.20, 0.25	0.32, 0.29	0.32, 0.32
WE_{Concat}	0.32, 0.29	0.36, 0.30	0.12, 0.14	0.36, 0.28	0.44, 0.36
WE'	0.65, 0.51	0.39, 0.40	0.39, 0.39	0.65, 0.51	0.39, 0.39
$WE_{DM'}$	0.70, 0.67	0.61, 0.63	0.39, 0.43	0.65, 0.63	0.52, 0.49
$WE_{Concat'}$	**0.70, 0.67**	0.65, 0.62	0.30, 0.31	0.65, 0.61	0.43, 0.46

Table 28: Houman Classification, Original vs. Refined Labels

The coherence-change index could be applied in this manner to a variety of combinations between the two labelling sets to guide and to improve classification's performance through multi-participant supervised iterations and it was useful in our context of Hafez poetry classification.

4.2.4 Classification Using Embedding Feature Experiments

We used word embedding as features (Mikolov et al., 2011), which was the basis of our model (Doc2Vec[40]) (Řehůřek and Sojka, 2010). Zhang and Lapata (2014) used word embedding in the poetry generation task and found it a powerful feature for capturing the semantic and context. To compare its impact, we kept the feature set and ML algorithm constant in each experiment. We used two Doc2Vec feature sets: Distributed Bag-of-Words (DBOW) and the Distributed Memory (DM), both separately and combined. Table 28 shows accuracy and F1 performance measures for different ML algorithms such as SMV, Regression, Decision-Tree, Neural-Net and Random Forest, abbreviated on the top row respectively. The last 3 rows in the table are after excluding inconsistencies, and the first 3 raws are only Houman's labels.

WE feature stands for DBOW, word embedding, and W_DM for that of Distributed Memory and W_{concat}, combined with the two feature sets (Mikolov et al., 2011). As we see

[40]We used *gensim* implementation.

in Table 28, the SVM algorithm in conjunction with the concatenated word-embedding features of CBOW and DM, plus exclusion of inconsistencies, lifted the accuracy of the automatic classification to 70% and the F1 score to 67%.

We compared Houman labels with those of a contemporary Hafez scholar, Raad. We not only introduced new effective features to automatically classify our Hafez corpus but were also able to find a new purpose for the experts' disagreements. We showed that by careful identification and exclusion of certain poems we could drastically improve the classification accuracy. The three top rows in Table 28 are before the exclusion and the bottom three raws are after the exclusion of such poems from the corpus. We also proved that SVM not only showed this phenomenon but also outperformed some other machine learning algorithms.

In other words, we measured the deterioration in Coherence; compared the effects of the inconsistent poems; and excluded the inconsistent instances if it improved coherence for both annotators equally. In Table 28, WE stands for before and WE' is after the inconsistent poems are excluded from training. Concatenated word-embedding features of CBOW and DM, plus exclusion of inconsistencies, lifted the accuracy of the automatic classification to 70% and the F1 score to 0.67. Inconsistency filtering lifted the accuracy of WE-SVM from 0.48 to 0.65 and F1-score from 0.31 to 0.51.

5

Semantics of Homothetic Clusters

In this chapter, we explain how we have created two clustering sets of semantic labels for the poems of Hafez (1315-1390), using unsupervised learning. We used clustering to generate new labels as an alternative to Houman's previously existing, hand-labeled, gold-standard classification of Hafez poems. We have cross-referenced, measured and analyzed the agreements of our clustering labels with Houman's chronological classes. Our features are based on word embeddings and are derived from topic modeling. We also introduced a new feature: similarity of similarities. We refer to this clustering with new similarity features as "homothetic". This transformation proved effective during our clustering experiments on the Hafez corpus. Homothety is a similarity transformation of certain attributes of vectors in Euclidean space that inspired my clustering algorithm. This approach produced distinct clusters, in the case of Hafez's small corpus of ghazals. Although all our experiments showed different clusters when compared with Houman's classes, we think they were valuable in their own right. Our clusters provided further insights and 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 for other similar literary research.

In his book, Houman (1938) partly hand-classified Hafez's poems, based on the semantic

attributes latent in the ghazals. His labeling has been the gold-standard of chronological classification for Hafez. In the previous chapter, we used it as training data for supervised learning to predict Houman's labels for the rest of the ghazals. In this chapter we use similar semantic features, but instead, we conducted unsupervised learning (clustering experiments) to create labels alternative to those of Houman.
Houman's classification was based on the premise that the artist's mindset and worldview changed throughout his lifetime and this change was reflected in his art, in this case, poetry. Houman wanted to hypothesize how time affected the meaning of Hafez's poems. We used machine learning to capture this chronologically changing worldview in the semantics of Hafez's poetry. For example, Houman believed that Hafez became more introverted with age. Houman explained in detail that these worldview characteristics and their interpretations were buried behind the surface meaning and in the semantic attributes of Hafez's highly indirect, multilayered and equivocal ghazals. Hafez's semantics were intertwined in the couplets and hemistichs but differently throughout his life.

5.1 Problem Statement

We hope that the chronological classification of Hafez's poetry will facilitate interpretation and demystify the depth of meaning in his majestic oeuvre. In this chapter, we use clustering as a semantic analysis tool to assist with literary investigations of Hafez's ghazals; that is to find out about the characteristics of the group they belong to. As a result, we have produced new unsupervised labeling visualizations for Hafez corpus[41]. We have also conducted what we refer to as *homothetic clustering* experiments, using similarity transformations as features, discussed in Section 5.2.4. We have performed semantic analysis by using topic terms, partly discussed in Section 5.4, using a topic modeling interactive visualization tool.
Although the fundamental question was to find out how consistent our semantic-featured clustering would be with Houman's chronological classification and to establish a verification experiment against Houman's labeling, we also set out to achieve the following objectives:

- Semantic Feature Engineering;
- K-Means Clustering (Automatic Labeling);

[41]Our Hafez corpus is available through in Ganjoor, Nosokhan and in Hafizonlove: http://www.hafizonlove.com/divan/index.htm & https://ganjoor.net/ & http://www.nosokhan.com/

- Similarity Feature Transformation as Homothetic Clustering;
- Multi-label Comparisons, Semantic Analysis and Visualization (Houman vs. clusterer).

By exploring Human labels in comparison with clustering results, we also wanted to see if our homothetic features could qualify our unsupervised method as a guided or quasi-semi-supervised labeling algorithm. Gieseke et al. (2012) optimized patterns in the data in the absence of labels and used SVM for classification.

5.2 Methodology

Our focus was to observe the performance and identify the semantic features that provided us with the best clustering results, measured by *Silhouette* (Kaufman and Rousseeuw, 2009). We were also interested in finding out which features produced results more consistent with Houman labels. To measure inter-annotator agreement we used *kappa* (Artstein and Poesio, 2008) and other measures (Viera et al., 2005). In all the experiments, we kept the clustering algorithm (K-Means) constant to isolate and focus on the effects of different features.

5.2.1 Preprocessing

We followed Asgari and Chappelier (2013) for our preprocessing steps, while being sensitive to Persian linguistic rules:

- Tokenization
- Normalization (This step scales the transformations to the unit norm and neutralizes the lengths of the vectors. This step sometimes improves performance.)
- Lemmatization
- Filtering (This steps gets rid of punctuations, stop-words and non-standard characters.)

In our preprocessing, we removed the stop-words and the tokens that occurred only once, as we did in our previous experiments. We built the dictionary of documents, every document being a ghazal. Then using the bag-of-words method, we set up and transformed the corpus into vector representations. We built the TF-IDF vectors. We

initialized LSI, LDA[42], 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 the HAZM[43] Python library for Persian pre-processing tasks, such as *lemmatization*.

5.2.2 Clustering Evaluation Indices

We followed metrics and clustering agreement techniques and scores[44] to measure our performance results in comparison with Houman's chronological labels. A perfect consistency mean identity or the value of one in the following measures.

- *Inertia*: Within-cluster sum of squared distances, which K-Means clustering tries to minimize; lower inertia is 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 direct correspondence between Houman classes and our clusters; in other words, the elements of the same class fall in the same cluster;
- *V Measure*: The harmonic mean of Homogeneity = HOM and Completeness = COM:

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

- *Adjusted Rand Index* (ARI): A similarity measure between clusters by pairwise comparisons of cluster and Houman class poems, E stands for *Expectation* in probability or weighted average of probabilities. Steinley (2004) used this index in cluster validation research: ARI is based on Rand Index but adjusted for chance.

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

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

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

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

- *Adjusted Mutual Information*: A symmetric measure of dependence between our cluster membership and the Houman class. Mutual Information (MI) is a measure of shared information between two clusterings U and V. H is the entropy. Vinh et al. (2010) normalized and adjusted MI for chance:

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

- *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 average mutual dissimilarities of objects in each clustering (Kaufman and Rousseeuw, 2009); they are aggregated intra-cluster and nearest-cluster distances of each poem to others.

- *Cohen's kappa* measures the inter-annotator agreement between two sets of labels, generated by classification or clustering .

5.2.3 Feature Engineering

We mapped poems into a vector space, using semantic transformations such as LDA (Topic Modeling) and Doc2Vec (Word Embedding). The variant of TF-IDF we used was based on logarithmically scaled frequencies of term i in document j in a corpus of D documents:
The LDA[45] implementation followed Hoffman et al. (2010); base code was found here[46]. We kept the default parameters when we initialized the LDA model. For the LDA-driven similarities, we only set the number of topics and passes to 5. We chose 5 to keep the topics intuitive for human review and it sufficiently satisfied our empirical purpose.
Doc2Vec[47] 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. We kept the default parameters provided by LDA implementation of gensim library.

5.2.4 Homothetic Features

The homothetic function is a positive, finite, continuous and strictly monotonic transformation of a homogenous function (Lau, 1969). Homothetic transformations are fre-

[45]https://radimrehurek.com/gensim/models/ldamulticore.html
[46]https://github.com/blei-lab/onlineldavb
[47]https://radimrehurek.com/gensim/models/doc2vec.html

quently used in transferring arguments amongst economic models (Christensen et al., 1975). To bring the Hafez poems to a vector space, we map each poem to a vector representation, using a mathematical function, using LDA driven cosine similarities. Our transformation is each poem's similarities. Intuitively, one could think of our algorithm 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, which we refer to as anchors. Anchor poems were chosen for semantic reasons to guide the clustering towards Houman's classes. For example, we chose anchors from Human's extremes or peripheries of each class. The criterion behind the choosing the anchors is based on maximization of the chronological distance. Using 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 vector to vector Euclidean distances (similarities), in a clustering procedure such as K-Means. Before passing the data to the K-Means algorithm, we transform the poems into their similarities to the anchors.

5.2.5 Homothetic Properties

Similarity-driven transformations are not necessarily linear and can enlarge distances. Similarity transformations also maintain *homothetic* properties, a monotonic transformation of a homogenous function for which the level sets (contour lines) are radial expansions (distances to origin) of one another. In Euclidean geometry, a homothety of factor k *dilates* distances between points $|k|$ times, in the target vector space. The associated risk of overfitting is higher with homothety. Because of the dialation power of homothetic features, the divergence rate is empirically much quicker. The properties of homothetic functions were proven by Simon and Blume (1994). If function v is monotonic, it is homothetic and reverse if, ($v = g \circ u$; x and y are vectors):

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

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

We want to investigate empirically, that the homothetic clustering procedure we use here, is effective to increase Silhouette score and is interpretable when used against our small

Data: Hafez Corpus
Result: labels
read corpus and anchor instances;
tokenize, remove stop-words and unique tokens;
normalize, lemmatize;
create *bag-of-words, TF-IDF*;
initialization of 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 the value of k clusters;
cluster (*Sim-Corpus*);
produce labels;

Algorithm 1: Homothetic Clustering, Sim^2

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

poetry corpus of Hafez. The average complexity of the homothetic clustering is 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 is the number of samples n, times the number of iterations i, times the number of clusters k:

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

5.3 Homothetic Clustering Experiments

In the first set of experiments, we used various semantic features for clustering. We then passed the vector representation of the labeled portion of the corpus to K-Means[48] for clustering ($k = 3, 6$). We chose 3 and 6 because we had done classifications for 3 and 6 of Houman classes. Then we compared the clustering labels with Houman labels. Table 29 shows the results. As we see, the Doc2Vec-PE 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 30, pure Persian *Embedding, (Doc2Vec-P)* showed the highest *Silhouette* value defined in Section 5.2.2, while adding English[49] to the corpus brought this measure a bit lower and still maintained second rank compared to all other features.

Houman (1938) selected a representative poem for each of his classes. We experimented both against three, maintaining the chronological order, and six Houman classes in Table 31. Since we selected the anchor poems from the corresponding Houman 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 number of clusters. The resulting

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

[49] English translation of the poems by Shahriari, 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 30: K-Means Performance
P=Persian, E=English

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

Table 31: Corpus Training Labels

6 Classes		3 Classes
Youth = 38	a	a'
After Youth = 25	b	
Maturity = 79	c	b'
Middle Age = 66	d	
Before Elderly = 28	e	c'
Elderly = 13	f	

either extreme periphery of the Houman's labeled poem classes, that is three from the earliest *Youth* class, and three from the latest period ranked in *Senectitude*. We referred to this experiment's feature set, Houman Extreme Picks (HEP). Or in case of the three classes HEP, we picked two extreme poems from either end of the class, and one central poem from class two, *Mid-age* (Houman ordered the poems chronologically). RND stands for random picks. We always ensured that the number of anchors matched the number of intended clusters: $(anchors = k = 3, 6)$, shown in the tables.

As we see in Table 32, HEP, HRP and RND maintain zero *Inertia* (within-cluster sum-of-squares) which is the optimal. This is an indication of favorable inner coherence of the clusters. HRP has about 3% as the highest *Homogeneity*, which was higher than that of the other two, Table 29. LDA had the highest *completeness*, while Doc2Vec-PE had the highest *AMI*. Both HRP and HEP models with similarity features also produced higher

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

Silhouette scores in clustering (Table 33) than the one achieved by the RND model, with word-embedding features. Only HRP showed slight resemblance to Houman's classes, as kappa values indicated in the same Table, although all indices are low but comparable. This means that Houman's selected poems, which he mentions in his book as their class representatives, in explaining his methodology, had a better homothetic guiding power than the actual extreme poems of his classified corpus, when we used them as anchors. We noticed a slight improvement in kappa, comparing HEP and HRP vs. RND which is based on random anchors. Kappa ranges indicate that closer to 1 there is agreement, closer to -1 disagreement and around 0 means that there is no conclusive information. We used random anchors (RND) which similarly did not produce sizable kappa yet was a bit worse that those of HEP and HRP.

The number of LDA topics in multiple K-Means runs affected the Silhouette score, but mostly converged at around 5 to 15 topics, depending on the feature set. To avoid local optima, it was also important to iterate through K-Means algorithm many times to attain an optimum Silhouette score while targeting the right number of LDA topics, to achieve the best possible clustering quality by trial and error. Our homothetic experiments achieved the best *Silhouette* scores with 6 LDA topics. In all homothetic and non-homothetic clustering experiments, the number of clusters $k = 6$ and $k = 3$, achieved the highest Silhouette scores, in their experiments group respectively, $k = anchors$. In the homothetic experiments, $k = 6$ clusters always produced both better kappa (comparing only when $k = cls$), 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. We refer to Poc2Vec-P as the challenger model because it was the best performing clusterer in the absence of homothetic features. The Spearman correlation was 0.86. It is noteworthy that the Cohen's linear and nonlinear *Kappa* were 0.58 and 0.43 respectively, between these two independent clusterers.

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 33: Sim^2 Performance, kappa with Houman classes

In this case, we did perform our Student's t-test, which 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. But inter-annotator agreement was a bit evident using *kappa* but that gives us very little information. 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.

5.4 Analysis and Discussion

We used the Persian part of the corpus for this section to demonstrate the semantic attributes and characteristics of our new sets of classes. Graphical overlay of the clusters and Houman classes did not show any significant overlap. Therefore, we do not perceive a chronological order for the segments.

5.4.1 Cycle of Words

More rigorous analysis should be done by literary scholars, such as deep interpretation, but as a sample of semi-automatic examination, we organized term frequencies in Figure 10 as follows. We counted the Houman-labeled poems in each cluster and calculated their percentages to decide the highest resemblance of each cluster or common number of poems with its closest Houman class. In the case of a tie, we did the same for the other clusters and then tracked back to maximize the overall resemblance by maximizing the completeness as much as possible, yet did not observe a pattern. HRP and HEP were constructed as explained in Section 5.3. Then we considered a cluster of terms, relevant to Houman's representative poems and his semantic constructs (Houman, 1938). For the Youth class (A), we chose three terms: Duplicity (*rıâ*), Sufi (*sufi*) and Abstemious (*zâhəd*). For Mid-age class (B), we chose Vision (*nazar*), Barmaid (*sâqi*), Knave (*rənd*). Finally for the Senectitude (C), we chose three representative terms of

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 10: Tracing Clusters of Terms

Expedient (*maslәhat*), Guru (*pɪr*), Pub (*meikade*). The terms are the top most frequent in each Houman class and heuristically viable to gauge clusters' semantic characteristics. Then we counted the frequency of the corresponding terms in each cluster, depending on the closest Houman class. Each cell in Figure 10 contains frequencies of its three terms respectively. There is no obvious or conclusive pattern to indicate a segment purely has more frequency of specific Houman-class-terms.

If we trace any effect of anchor meaning in the final homothetic clustering result, we observe that HRP has a slightly stronger resemblance with the Houman classes as it was also measured by higher homogeneity and completeness in Section 5.3. Both HEP and HRP showed the better overall balanced distribution in terms of the size of each cluster compared to Doc2Vec-P, which was also reflected in the higher Silhouette score from Section 5.3. Although both HEP and HRP showed a stronger correlation with Houman classes than Doc2Vec did, the HEP and HRP had 0.58 *kappa* and 0.86 correlation. HEP was also stronger in distinguishing between class A and C because we had purposely selected its original anchor poems 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 semantic properties of the transformations of the *anchored* poems.

5.4.2 Results

- Doc2Vec-P word embedding scored higher coherence[50] and Silhouette than other non-homothetic features used in the automatic clustering of Hafez's poems;
- We created two new sets of automatic labels for the Hafez corpus, by Doc2Vec as challenger and Sim^2 as champion clusterers, which had 0.58 *kappa* and 0.86 correlations but had statistically insignificant resemblance with the Houman labels, 0.034 kappa at best (HRP-6cls); There was no significant observable pattern among clusters and Houman classes to show.[51]
- Sim^2 did not fully qualify as a quasi-semi-supervised[52] algorithm, given the low 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 significant consistency with Houman's classification, but provided with new semantic perspectives to Hafez studies;
- Semantic evaluations and visualizations helped validate the clustering results, using random poems; the LDAVis library was used to depict relevant topic-terms by clusters; examples are shown in the appendix.
- Visualizations in conjunction with homothetic clustering could be used to analyze the semantic properties of Hafez poems, even with small corpora such as ours.

Inspired by Houman's semantic approach, one can replicate and apply our poetry clustering arapproach to other poetic texts, as a means of assisting and enabling literary research and scholarly analysis of poetic text by clustering only if possible and when the scarce data conditions were similar to ours. In this chapter, we provided with the blueprint of an effective clustering of Hafez poems. Our guide is with reference to Houman's order of poems, which is based on Ghazvini's copy, which is an old and reliable printing edition of Hafez's poems, organized alphabetically.

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

[51] The Sim^2 clusters are available in Section 7.4.

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

5.5 Conclusion

To support capturing semantic attributes of Hafez's poetry, Houman's proposed a chronological and semantic classification, unique up to now, assuming 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 can help to produce different clusters of poems, to complement their scholarly hand-labelled version. We introduced the similarity-based features to build our better performing 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 which were giving us slightly higher kappa compared to that of the random anchors. In the 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 from (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 comparable to Houman's classifications. They are considered semantically comparable because we used very similar semantic features but only removed the gold-standard labels.

We applied our homothetic features that proved the most effective in our clustering, to label the whole Hafez corpus as parallel labelling to Houman's. We investigated semantic differences, using both labels while comparing and tracing the consistencies through visualizations. We performed heuristic and empirical semantic analysis, and tried to refine and guide 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 and analyze Hafez poetry.

6

Semantic Results, Visualization and Analysis

6.1 Scholarly Views of Hafez's Poetry

In this chapter, we used the inter-annotator agreement to examine the validity of Houman classifications against another contemporary scholarly chronological classification of Hafez, by Raad (2019). According to Roland Barthes in "The Death of the Author" (1967) each scholarly perspective or interpretation of Hafez, almost 100 years apart, has its own independent unique stance and value. Regardless of the invaluable independence of the two scholarly perspectives of Hafez, not only did we want to compare the inter-annotator agreement between them to generate insights but in the reverse direction and as a side benefit, we also wanted to see whether their labelling inconsistencies could guide us to go back and improve the automatic classification. In the end, we realized the two perspectives need not necessarily be perceived as contradictory but overlayed to deepen our semantic awareness.

6.1.1 Houman's Perspective on Hafez

This section reiterates some points from section 3.1.2. The basic question was which version of Hafez's line of ghazal was correct where there was any ambiguity or difference in a word, diacritic etc. Houman (1938) presents Hafez as an extraordinary poet and philosopher whose poetry includes mysticism, resentment against pretentious clergy, demeaning towards dialectic philosophy while praising love, largess, tolerance, liberty and joy, all encapsulated under *Rend* attributes. Houman firmly sees and encourages the deep interrelation of the ghazal lines and, as many scholars agree, the first line often carries the main topic of the ghazal. Houman was a pioneer to establish and maintain that understanding of Hafez based on cohesive facts was possible; and although there was not very much supporting documentation available to him other than different versions of Hafez's poems, Houman showed that one could still understand them but by setting aside subjectivity, personal tastes and presumptions as much as possible and rely more on pragmatic rigour and unbiased logical analysis of the whole corpus.

Houman studied the genealogy and the meaning of the symbols and expressions Hafez used in the poems to categorize Hafez's thoughts into main and secondary concepts. According to Houman, some of the main ideas circles around terms such as 'knave' *[rend]*, 'love', 'wine'; some of Hafez's secondary thoughts are related to 'destiny' and being 'dervish'. Through the logical methodology and analysis, while considering the supporting geopolitical and historical events at Hafez's time, plus attention to psychological and philosophical properties, Houman could match Hafez's poetry to an evolutionary worldview and accordingly mapped Hafez's poems into an ordered sequence of time slots of his lifetime.

Houman perceived Hafez, as a philosopher-poet who questioned the universe. Hafez did not seem to believe in the plots set forward by the religious documents. Houman brought references from the ghazals that supported his claims. For example, the last line of a ghazal reads: 'Hafez, our existence is a mystery, solving it is but all imaginary'.
Houman's classification of the ghazals is based on the natural evolution of the poet, and therefore Houman analyzes this evolutionary process starting from questioning Hafez's philosophy of being and the way Hafez viewed the origin of the universe; as a result, Houman for example, can explain why Hafez was in awe.

Houman follows the semantics of expressions through the previous literature and cultural history and draws such conclusions that relate the perceived "wine drinking" with demeaning intentions towards clergy and religious ideologies; he brings examples of other

giants of Persian literature that used similar expressions to disregard the dialectic and abstract logic. Houman interprets the metaphor of wine drinking and its praise by Hafez as a symbol of dislike towards the misguided knowledge that was solely based on a logic that had no legitimate sense or rational philosophical context as opposed to mystical intuitions and metaphysical inspirations.

Houman also did psychoanalysis on the bipolar emotional reasoning behind Hafez's preferences; for example, Houman mentioned love against knowledge, praising the unworthiness of the world against notions of power and affluence. Houman presents Hafez's attraction toward turning his back to social rank by honouring doubt. Houman mentions the praising of love against imaginary hopes promoted by religious dogma as valuable clues and indications in understanding Hafez.

The meaning of 'Rend' that is 'knave' is somebody who is against the asceticism and the ascetic. Houman referred to examples of such contexts that had been brought forward in works of Khayam and Sanaii to support this use of "knave" prior to Hafez. According to Houman, Hafez's attitude is reflected and demonstrated in his ghazals. For example, turning his back to affluence and materialism is rooted in Hafez's perception that pretenses and charlatanism came from the weakness of character, which in turn was a roadblock to human greatness of self. Therefore one should always avoid attractive superficial and decorative earthly material and instead thrive for simplicity, and search for cleanliness in psyche and constantly look for unbiased truth, shown in Figure 11.

Hafez also has three distinct types of references to love in his poetry, according to Houman: Freudian, Platonic, and mystic love depending on the context and semantic factors. These love types are contradictory and exclusive of one another and therefore hardly happen in the same line, but they existed throughout the life of Hafez and therefore these concepts existed throughout his poetry. Hafez transitions from one type to another according to the chronological context, confirming that he even maintained the detached knave's attitude towards love itself. Houman brings much supportive evidence from the ghazals, in which Hafez shows love towards righteousness or God and refers to humans as realizations of virtue. Wine also closely relates to the meaning of love; wine essentially revives and makes love come to life.

Hafez has references to destiny as cause and effect and sometimes as an unchangeable and predefined event. Houman maps these profound changes to his maturity growth line. Similarly, Hafez has references to dervish attributes to reference the freedom and sometimes to mean withdrawal, abstinence and contentment. Houman uses these con-

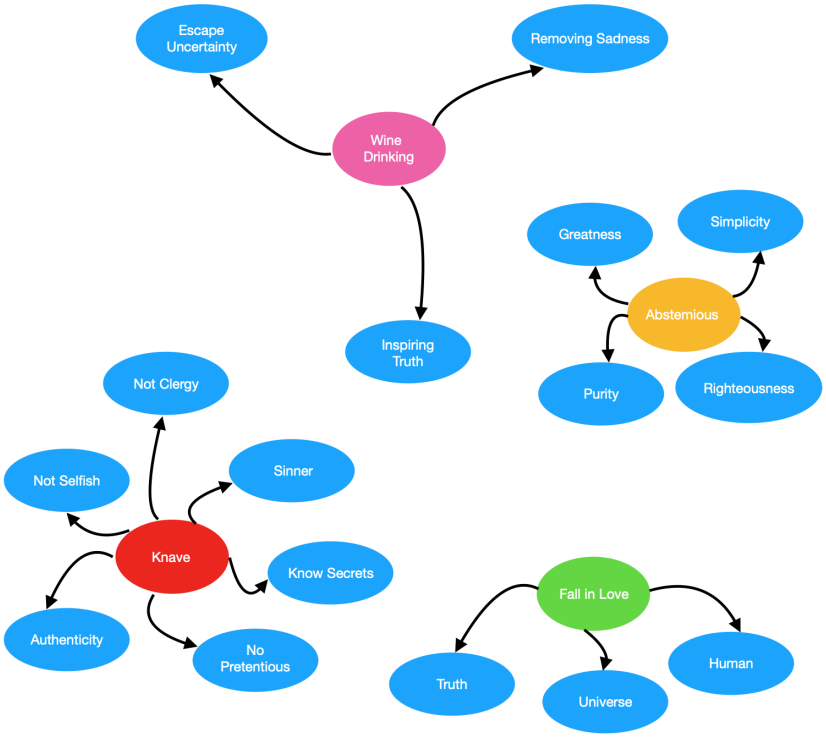


Figure 11: Houman's ontology of Hafez's work

cepts and analytical results to map the poems to the chronological evolution of Hafez's character and ideology and his maturity growth.

6.1.2 Raad's Perspective of Hafez

Raad (2019) viewed and analyzed Hafez from four figurative dimensions or elemental perspectives:

- Virtue;
- Politics;
- Technique;
- Time periods.

In terms of *virtue*, Hafez elaborates on ideas from philosophy, sophism, religion, and anthropology. When it came to *politics*, Hafez was sensitive to the improvements of social affairs, security, he discussed and cared for political figures and their effects on social classes. *Technically*, his poetry was full of linguistic innovations, cerebral imagery,

and discerning characterizations, meanwhile coming across as apt and ingenious. Raad (2019), very similar to Houman, believed in the effects of *Time-Periods* as an evolutionary or temporal dimension in the pursuit of dissecting Hafez's poetry. However, Raad (2019) distinguished and divided Hafez's chronological life-periods only into four sections as opposed to six, named by their associated two leading political figures:

- Youth, before *Amir Mobarezzeddin*;
- *Amir Mobarezzeddin*;
- *Shāh Shojā*;
- After *Shāh Shojā*.

6.1.3 Analysis and PCA Visualizations

In this section, we discuss the rationale behind the classifier predictions. We developed the graphical representation based on the LDA-PCA model, using predictions and test data.

In all graphs, the x and y-axis are PCA dimensions that were reduced from the LDA vector space to two dimensions, to acquire a 2D representation. PCA is a feature extraction technique that reduces the dimensions using matrix manipulations while maintaining the important features and information as much as possible. As mentioned earlier, Houman classified poems according to their meaning, which corresponds with the maturity development of Hafez's philosophy and spiritual path. According to Houman, in the first period of his life, Hafez did not pay much attention to the conditions of his time and was instead more playful, *"knave"* and *"gaze"*. In the second period, the hypocritical behaviour of the powerful clergy was more evident, and his reactions to this became more sympathetic to human suffering, moving him toward mysticism.

In the third period of his life, Hafez was more appreciative of happiness particularly when drinking wine, and he developed a sense of freedom and joy, and all is well and cheerful.

According to Houman, in the three periods above the effects of external and internal factors were balanced and equally important. In period four, however, Hafez's endeavours, effects and attention to internal affairs become much stronger; he again paid less attention to the conditions and surroundings and focused more on his internal path and godly inspirations. And in period five, we see strong attention and obligation toward mysticism, which reveals elements of his unique, deep and eloquent love and passion

toward life. There was a transformation in Hafez and his feelings about love, wine and *"knave"*. Finally, in the last period of his life became more introverted, and busier with his internal passionate spiritual love and human inspiration.

In addition to this reasoning, in this section, we select and employ appropriate ML tools as analytical methods to generate insights and present evidence of how topic modelling and PCA results correspond with Houman's classification concepts. We found this approach extremely useful, and believe the methodology and results in this section are very educational. We would recommend studying our results to anyone interested in understanding Hafez through his ghazals. Houman's chronological labels help us understand Hafez's poetry better.

We used the LDA-PCA method to create visual artifacts and analyzed parts of the Hafez corpus. This provided intuition and insight into how the ghazals relate and which semantic factors were contributing features to the classification.

The output of our visualization tool was the top extracted terms that were representative of each class; we call this a cluster of terms. These clusters correspond with the number of topics, and we will see that the darker areas are distinctive topics. Our tool lets the user expand and magnify the clusters, and observe the frequent Dirichlet terms that form them. The visualization also provides a graph or network of the weighted Euclidean transformation of the highest probable terms, based on the Dirichlet distribution of the topic for the class. It is evident that when the topics relate, that is when their weighted distance is above a certain threshold, there is an observable edge between the two topics. In addition, the long or short distances between topics correlate with the stronger or weaker relations, respectively. In this case, stronger relations between topics indicate a shorter distance, and more appropriate semantics.

Hafez's first period cluster of terms across all topics relate to *Hair* and *Hand* and *Heart*, *Flower* and *Cup*, meaning they could represent youthful enjoyment. And we see that the second period cluster of top terms in all topics correspond to the beginning of his mystical endeavours, due to the connotations of *Sadness* and *Love, Dust, Life* and *Valley of Heart*. Interestingly, our model picks *Happy* for the third section or period of the corpus, which is consistent with Houman's classification logic that specifically defines happiness. In period four, the notions of *Universe* and *Disability* come into play, and *Love, Heart, Cup* and *Wine* are still present. We observed this as indication of stronger internal and philosophical inspirations, as Houman suggests.

We see that the term *Heart* is in all the periods, *Sadness* is only in the first two, *Hap-*

piness is in period three and onward and *Sadness* is not in any period after that. *Wine* starts in period three, and remains until the end, but in different cycles of terms and context. According to Houman, *Love* is common in all periods, though with different intentions and less bold in the first and last.

6.2 Main Topic Terms of Class One: Youth

The Youth class has the following cluster of terms. Topic numbers are generated automatically by the *gensim* library, and topic terms correspond with topic numbers in the graphical representations in the following figures. Some words in the translations of topic terms look like stop-words; they were not stop-words in the ghazals but were linked to other words.

0. Vision *nazar*, Connected *vasl*, Unable *nâtavan*, Complain *ʃekâyat*, Your Sorrow *qamat*, A Heart *deli*, Glass *ʃɪʃə*, Repentance *tobæ*, Universe *jahan* and Hand *dast*.

1. Flower *gol*, Reminiscence *bovad yâd*, Airy *havâi*, Solution *tadbɪr*, Jam *jam*, Wine *məɪ*, Guru *pɪr* and Hand *dast*.

2. Sorrow *qam*, Blood *xʊn*, Wine *mɛy*, Full *por* and To Be Me *bâʃam*.

3. Arch *tâq*, Gem *laæl*, Because *bahrɛ*, Face *didɛ*, Speech *ʃirin-soxan*, Limit *hadd*, Business *kâr*, No Hint *nemibinam-neʃân* and Ruined *xarâb*.

4. Secret *sərr*, Destiny *qadar*, Say *gʊ*, Cup *jâm*, Know *dânɪ*, Friends *ʏârân*, Came *âmad*, Dawn *sahar* and Life jân.

5. Break *beʃkan-bɛ*, Title *maqâm*, Life *jân*, Thousands *ɦezârân*, Loose *sost*, Candle *ʃamæ*, My Heart *delam*, Love *əʃq* and Downhill *naʃib*.

As indicated in the Figure 12, topics 1, 2 and 5 are further from each other, and from the other three topics, assuming the geometric properties reflect interrelations, while in the network or weighted-Euclidean-distance Figure 14, topics 1 and 3 are not related to others in the graph, and topics 0, 2, 4 and 5 are related in that order or linked in that sequence. This indicates how the term characteristics of the topics interrelate. In this case, we are more likely to see topics 1 and 3 in a ghazal, but the cycle of words in topics 1 and 2 are not expected to be seen in the same ghazal. We can also observe the contrast between topics 1 and 2; that is, topic 1 is obviously unrelated and far from topic 2. The PCA visualization is maintaining a dynamic spacial relationship as distance, so

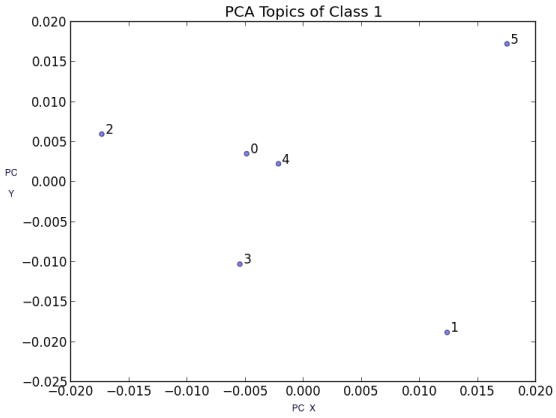


Figure 12: LDA Topics for the class Youth

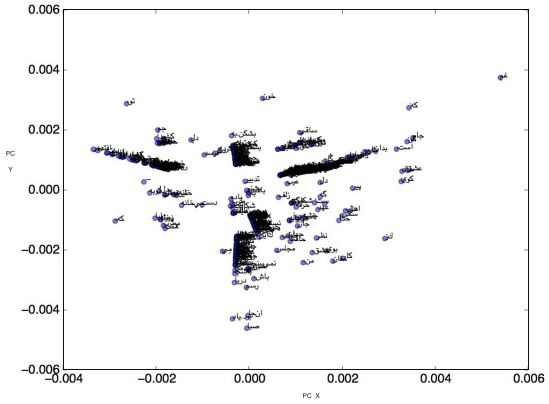


Figure 13: LDA Word Clusters for the class Youth

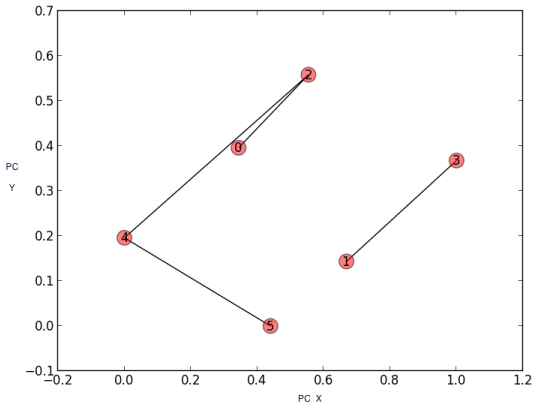


Figure 14: LDA Topics; Graph Relations for the class Youth

the positivity or negativity of the measures on the axis is not necessarily telling us any specific information. Figure 13 is a word-cloud for class Youth, composed of Persian words.

6.2.1 Analysis of Poems: Class Youth

Here, we examine a poem that Houman classified as belonging to the Youth period of Hafez's life, and identify the elements and circles of words in it. The first line of the ghazal 48 is this:
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.
Examining this ghazal shows that the terms *Glass*, *Heart* and *Sorrow* correspond with the topic 0. *Sorrow* also belongs to topic 2, *Is* occurs twice and *Be* five times. Interestingly, the network Figure 14 shows that there is relationship between topics 0 and 2. Elements of topics 1 and 5 are depicted far from topics 2 and 0 in the PCA chart, and thus are not present. Overall, the elements and genre of the ghazal are consistent with the concepts depicted by the word-cycles and topic-charts of this class. These are system-generated topic terms for this particular Houman class. The gist of the poem is about the youthful fever of love, when traveller, referred to himself, runs into a stranger and takes her hand, and they drink a glass of old wine together. He is looking for a missing joy and a lost love, although love is painful, he is willingly embracing the natural harshness of the beloved.

6.2.2 Main Topic Terms of Class Two: 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* and Concern *kâr*.
1. Vision *nazar*, Life *jân*, Return *bɑz*, Universe *jahan*, Cleanliness *taharat*, Secret *serr* and Is *ast*.
2. Hafez *hâfez*, Heart *del*, Soleiman *soleimân*, Virtue *honar*, Word *soxan*, Distressed *pariʃân*, See *bin*, Candle *ʃamæ* and Vision *nazar*.
3. Went *raft*, Return *bâz*, Not Remain *namânad*, Flower *gol*, You *to*, Sweetheart *yâr*, Harm *balâ* and Sympathy *deli*.
4. Envy *hasrat*, Said *goftâ*, Dust *xâk*, This way *kɛ-ɪn*, Cup *jâm*, Palate *kâm*, come I said *âyad-goftam* and Come *biâ*.
5. I want *xâham*, Has Left *nahâdæ*, Cannot *natavân*, Wrong *qalat*, Eye *ɕaʃm*, Contract *ahd*, Is-Not *nist* and Wine *məy*.

6.2.3 Analysis of Poems: Class Maturity

An example of analysis for this section is ghazal 206 of Houman's classification period 4. The first line of the ghazal is: *sɑlhɑ dafter mɑ dar geroye sahbɑ bʊd - ronaghe meikade az darso daɑye mɑ 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

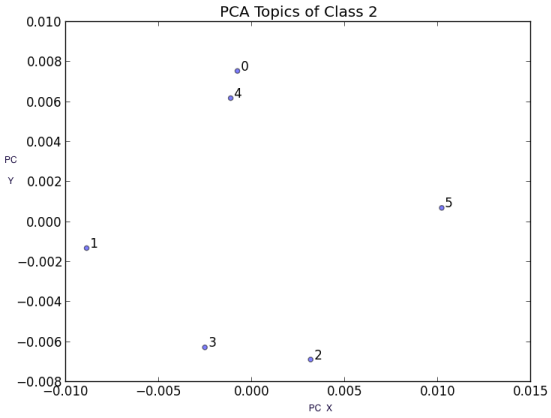


Figure 15: LDA Topics for the Class Maturity

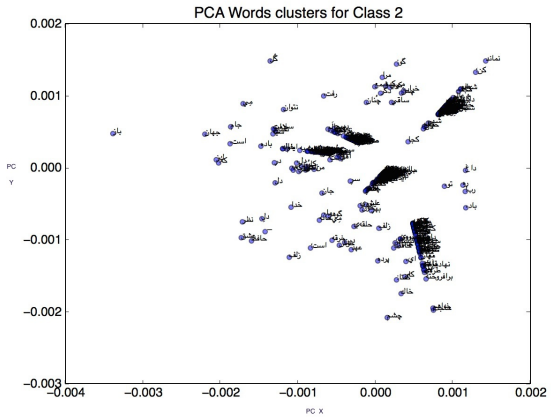


Figure 16: LDA Word Clusters for the Class Maturity

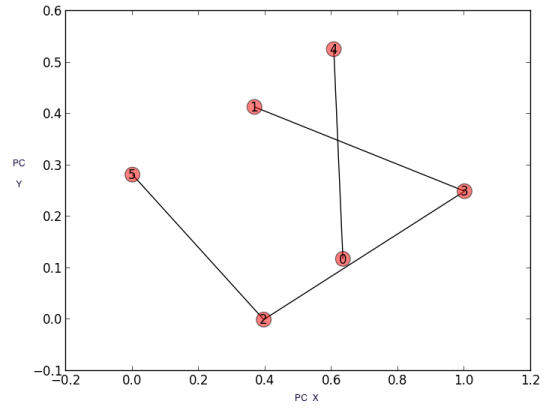


Figure 17: LDA Topic, Graph Relations for the Class Maturity

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 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.
In Figure 15 we see that the highest number of terms in the term cluster belong to topics 0 and 4: The term *That* occurs five times and twice in a slightly different form, and *Heart* and *Said* are common in topics 0 and 4. The system shows this relation in both the topics chart and network distance relation charts. We also observe the terms *Vision* and *Universe* from topic 1, See from topic 2 and *Flower* from topic 3, all occurring once. This relation is depicted in the network Figure 17. Figure 16 is a word-cloud, Persian words, for the class maturity. The automatically generated topic terms and their unique graphical depictions help us better figure out the semantic properties that Houman perceived for the class.

6.3 Main Topic Terms of Class Three: Elderly

The Elderly Class has the following clusters of terms:

0. Prescription *davâ*, Universe *donyâ*, Does it *bekonad-ze*, Wonder *ajab*, Happy *xoʃ*, Kindness *məhr*, Cup *jâm*, Is *bovad*, Veil *hejâb* and Free *rahâ*.

1. Life *jân*, Song *âvâz*, Scream *faryâd*, All *ʃamæ*, In *andar*, Nightingale *bolbol*, Universe *jahân* and Let it become *ʃavad*.

2. Full *por*, Sadness *qam*, Became *bəʃod*, Witness *ʃâhed* and 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* .

4. Fell off *oftâd-az*, Fell *oftâd*, My heart *delam*, Blood *xʊn*, Does from *konad-zə*, Hand *dast*, Universe *jahân*, Love *əʃq*, Familiar *ahl* and Smell *bʊyə*.

5. Better *behtar*, Wisdom *aql*, Turn *nɔbat*, Drink *bâdeh*, Within *andar*, Wine *mɔy*.

6.3.1 Analysis of Poems: Class Before Elderly

We randomly selected the ghazal 241, which Houman classified in the last class. It starts with the line: *har chand piro khaste delo natavan shodam - har gah ke yade 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 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 agrees with Houman's descriptions of the attributes of this class, as it describes a very introverted and sad poet who has few connections with the natural world, and specific mentions of Hafez referring to himself as old. Now we analyze how our developed cluster-of-terms played out in this case.

Although we can see the sporadic presence of nearly all topic terms, except those of topic 6, it is evident that topic 2 is dominant when we observe and identify the associated clusters of terms of this group: such as *That*, (Thrice), *Sad* and *Wine*. Topic 1 is next, with

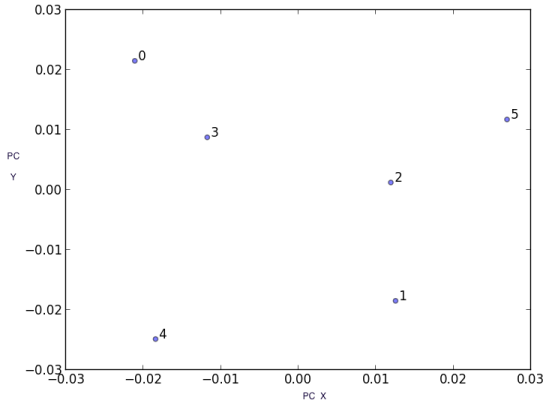


Figure 18: LDA Topics for the Class Elderly

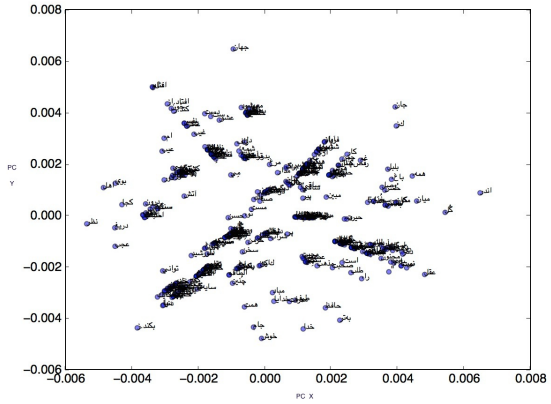


Figure 19: LDA Word Clusters for the Class Elderly

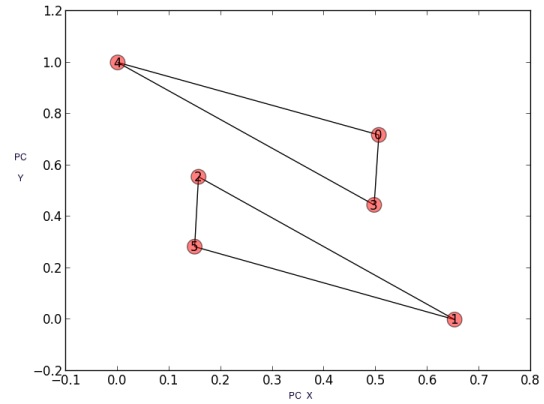


Figure 20: LDA Topic, Graph Relations for the Class Elderly

the terms *Nightingale*, *Universe* and *That-From*. The graphs show the geometric properties of topic terms in a 2D space for each Houman class to help us observe the interplay of the topics of that particular class. Our visualizations are intended to provide clues and therefore findings are based on anecdotal evidence.

The word clusters Figure 19 of this class have more concentrated clusters than the other two, which might explain the higher overlap of topic terms. Notably, since we chose this poem, which is deeply in the Class Elderly, topics 0, 4 and 5 which are farthest in Figure 19 also present partial terms. We see the terms *My Heart* and *Universe* of topic 4, *Cup* of topic 0 and *Wine* and *Is* of topic 5. *Wine* also overlaps with topic 2, and the term *You* of topic 3 appears three times.

The interesting symmetric nature of relation network Figure 20 for this class is consistent with our observations, in that there is a strong presence of a cluster of terms 1, 2 and 5 and weaker presence of topics 0, 3 and 4; though the term *Universe* is common. If we exclude *Universe* from both sub-graph terms, network Figure 20 shows there are only three distinct terms in the weaker group (i.e. 0,3 and 4) than in the other sub-graph (i.e. 1, 2 and 5) with the presence of term 7. Figure 19 is a word-cloud, Persian words, for class Elderly.

6.3.2 Predictions Validation and Analysis

In this section, we look at three different predictions to determine if they have the attributes of their class as Houman outlined. The predictions discussed here are from our champion model, which is explained in section 4.1.17.

Although our two-phased model for three classes predicted that the unlabeled data fell in two classes most in Maturity and some in Post-Maturity, according to our champion model of all six classes all predictions for unlabeled data fell into three classes: one, three and four, in proportions of 51%, 18% and 31% respectively. Those are the results for the ghazals which Houman left unclassfied. Classes a, b, c, d, e, f correspond with Houman classes 1 to 6 respectively. The champion model's confusion matrix is shown in Table 23, in Section 4.1.17.

Figures 21, 22 and 23 are network signatures of amalgamated Houman classes of *Youth, Midage and Senectitude*.

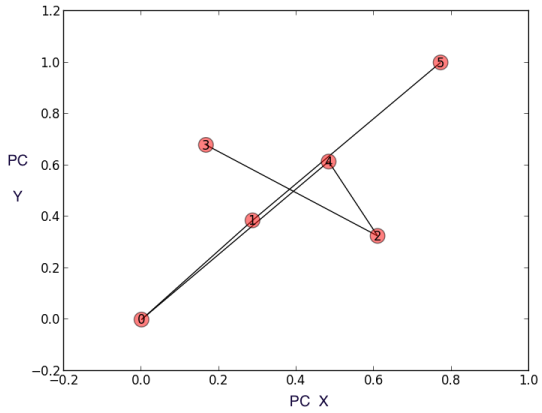


Figure 21: Class a', Topics Network

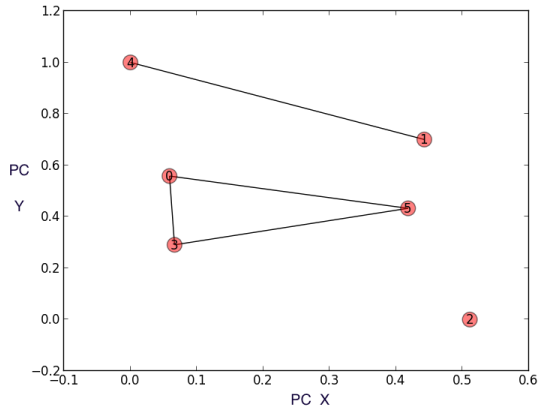


Figure 22: Class b', Topics Network

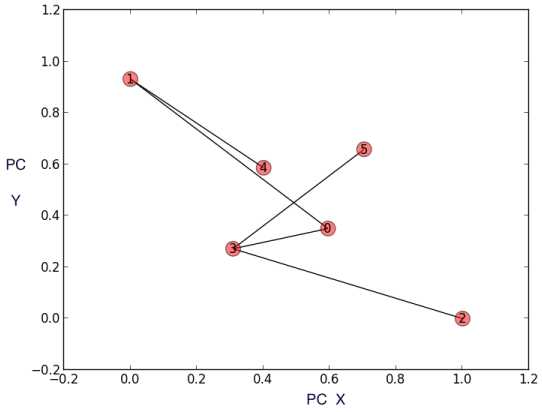


Figure 23: Class c', Topics Network

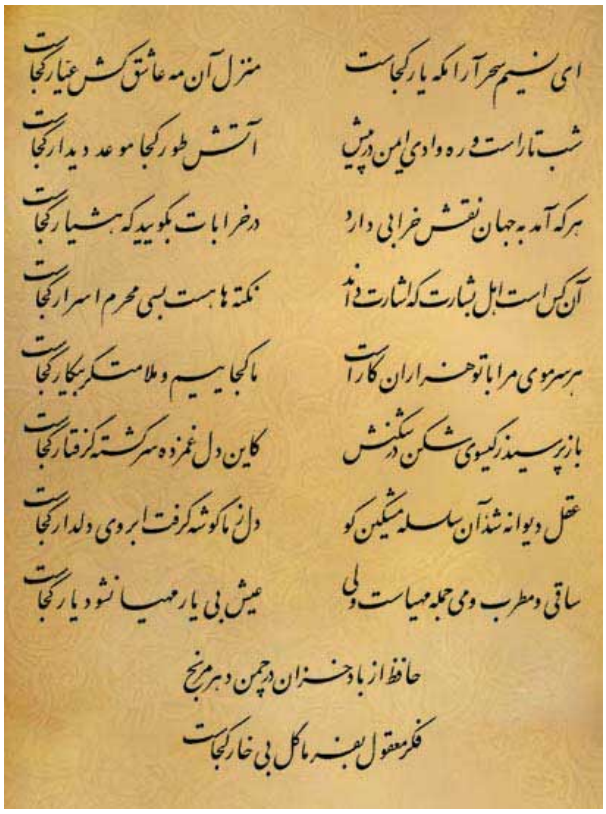


Figure 24: Ghazal from Class Youth

6.3.3 Poem Example One

We now assess a ghazal in Figure 24 that our model has classified in Class 1, Youth: *O fragrant morning breeze! The Beloved's rest-place is where?*
The dwelling of that Moon, Lover-slayer, Sorcerer, is where?
Dark is the night; and in front, the path of the Valley of Aiman:
The fire of Toor where? The time and the place of promise of beholding is where?
Whoever came to this world hath the mark of ruin:
In the tavern, ask ye saying: "The sensible one is where?"
One of glad tidings is he who knoweth the sign:
Many are the subtleties. The confidant of mysteries is where?
Every hair-tip of mine hath a thousand bits of work with Thee:
We, are where? And, the reproacher, void of work, is where?
Reason hath become distraught: that musky tress, where?
From us, the heart hath taken the corner: the eye-brow of the heart-possessor – is where?
The cup, and the minstrel, and the rose, all are ready.

But, ease without the Beloved is not attainable. The Beloved is where?
Hafez! grieve not of the autumn wind in the sward of the world:
Exercise reasonable thought. The rose without the thorn is where?

As we can see, the surface meaning and theme are about the beloved thriving. An extreme earthly or heavenly impulse that is a great sense of sensual passion for the beloved is throughout this poem. Hafez is searching for a somewhat painful aspiration for his beloved, and she is the only one Hafez refers to as his confidant. Thus, we observe the elements of the Youth Class that carry the weight of cravings.

If we only consider the cycle of high probability terms used earlier in the visualizations, without considering the inter-topic similarities, this poem would have fallen strongly into Class b'. This is due to the overall number of ten high probability terms with six topics in one of the runs. There are as many as 37 instances of existing terms in the poem for Class b', whereas Class a' and c' have 13 and 12, respectively. This alignment explains that when high LDA probability-based features are the driving force, the prediction classifies this poem as Class b'; it is then classified as Class c (Maturity) in the second and more granular modelling stage of Section 6.6. However, when the similarity factors are the driving force, the champion model classifies this poem as Class 1 (Youth) in Section 6.8. We first extracted the cluster of words that are characteristic of this poem, with reference to the class with the strongest topic term probabilities; in this case, it is Class b'. Then, using the network visualization technique, we depicted simplified similarities of six topics structure; that is, assisted human intuition by overlaying a distance-based network of topics. The poem has many elements of class a' such as *beloved, hair, eyebrows, cup-bearer, meadow.*

The following are the high probability topic terms for this poem, which we refer to as the cycle of words, that characterize the poem with reference to its predicted class by Stage 2 of our multi-staged challenger model, as explained in Section 6.6. These words are the top probable LDA terms from Class b' in the poem. This effect occurs most frequently with this class compared to Classes a' and c':

gol flower, *guʃe* corner, *kojâ* where, *dəl* heart, *âtash* fire, *marâ* me, *məy* wine, *bâz* again and *jahân* universe

We next assessed the distance networks, which are ordered left to right, top to bottom in Class a', b' and c', to determine if the poem's network of the topic distance structure (bottom-right) most resembles the one our champion classifier predicted: Class 1 (Youth):

We see that the last image most resembles the first; the intuitive and rather heuristic

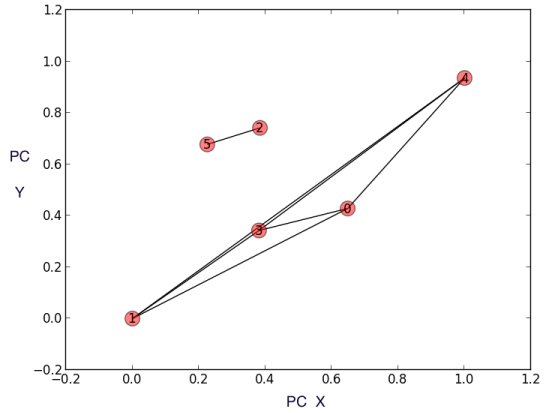


Figure 25: Poem One's Topics Network

criteria are based on the similarity in distance and connections of the topics. that is, there are two further topics that relate to each other!

Both human judges[53] voted that this poem truly belongs to Class 1, Youth. Their reasons were similarly: the joyful energy of the poem and lack of the other elements of later years such as attacking the clergy or mysticism or deeper worldviews and insights.

6.3.4 Poem Example Two

Both our champion model and the two-phased mode classified the next poem shown in Figure 26 as belonging to the Maturity Class.

Those of lily perfume cause grief's dust to sit when they sit:
Patience from the heart, those of Angel-face take when they strive.
To the saddle-strap of tyranny, hearts they bind when they bind:
From the ambergris be perfumed tress, souls they scatter, when they scatter.
In a life-time, with us a moment, they rise, when they sit,
In the heart, the plant of desire they plant, when they rise up.
The tear of the corner-takers they find, when they find:
From the love of morning-risers, the face they turn not, if they know.
From my eye, the pomegranate-like ruby they rain, when they laugh:
From my face, the hidden mystery, they read, when they look.
The one who thought that the remedy for lover is simple:
Out of sight of those sages who consider treatment, be.

[53]Hafez experts asked to classify independently are Mr. Mehran Rahgozar (Expert1) and Mr. Mehran Raad (Expert2).

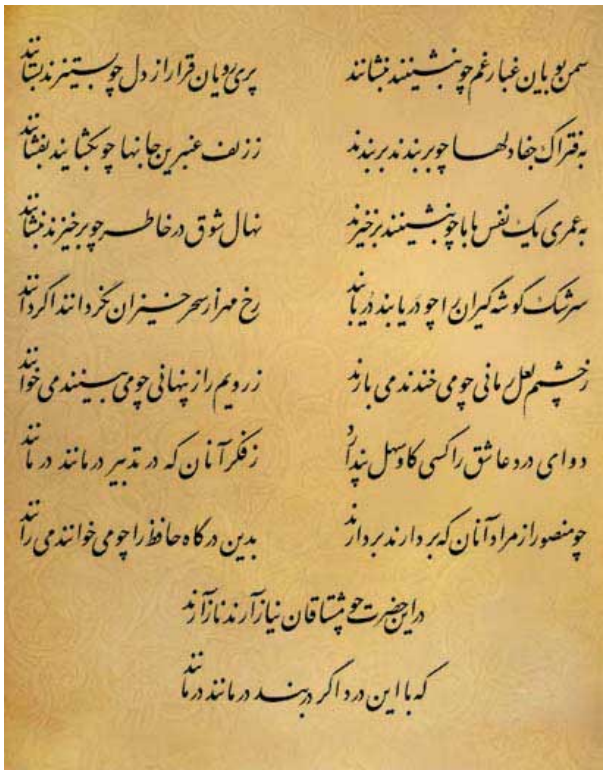


Figure 26: Ghazal from Class Maturity

Those who like Mansur are on the gibbet, take up that desire of remedy:
To this court, they call Hafez when they cause him to die.
In that presence, the desirous ones bring grace, when they bring supplication:
For, if in thought of remedy they are, distressed with this pain, they are.
The following are similar to the previous cluster of words example: *dəl* heart, *omr* life, *çaʃm* eye, *marâ* me and *jân* life.
We then assessed the distance networks that are ordered left to right, top to bottom in Class a', b' and c', to determine if this poem's network topic distance structure (bottom-right) most resembles the one our champion classifier predicted: Class 3 (Maturity).

As we see, the last image most resembles the top-right Class b', as there are two isolated topics that relate to one another and one that is further than these three!
Our expert human judges agreed that this poem belongs in Houman's Class 3, Maturity! Both expert judges reason that the poem has the ultimate style of mature Hafez and also has strong expressions of freedom that indicate Hafez has passed the sufism and narrower perspectives to life.

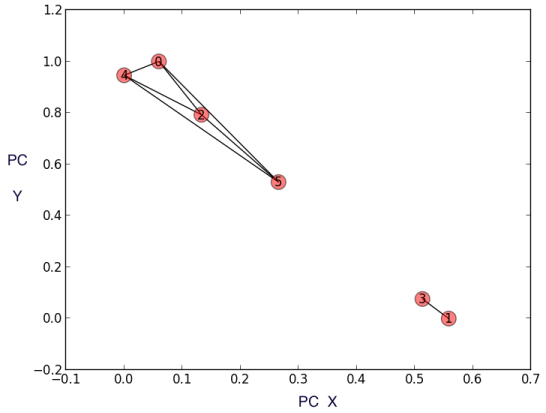


Figure 27: Poem Two's Topics Network

6.3.5 Poem Example Three

Our champion classifier classified the next poem shown in Figure 28 as Mid-Age as did the two-phased model.

The violet is vexed with envy of thy musk-scented tresses; at thy heart-rejoicing smiling the rose-bud rendeth its leaves.

O my perfume-exhaling rose, consume not thine own nightingale, who with heartfelt sincerity prayeth for thee night after night!

Behold the might of Love! how, in pomp and splendour, he dared, beggar though he be, to break off a fragment of the crown of royalty.

I, whom the breath of angels made sad, can for thy sake endure the quarrels of the world.

To love thee is the destiny inscribed on my forehead; the dust of thy threshold is my Paradise, thy radiant cheek my nature, to pleasure thee my repose.

The rags of the saint, and the goblet of wine, although they do not harmonize well, I have blended into one, because of thee.

Love, like unto the beggar, still concealth treasure in his sleeve; and soon he who was thy suppliant will be exalted to sovereignty.

The resting-place of thy form is my throne and altar: O my queen, do not abandon thy place.

This bewilderment of wine, and this delirium of love, will not depart from my head until I abase it, full of desire, in the dust at the door of thy dwelling.

Thy cheek is like a fair meadow, especially when, in the lovely spring, Hafiz, sweet of speech, is thy nightingale.

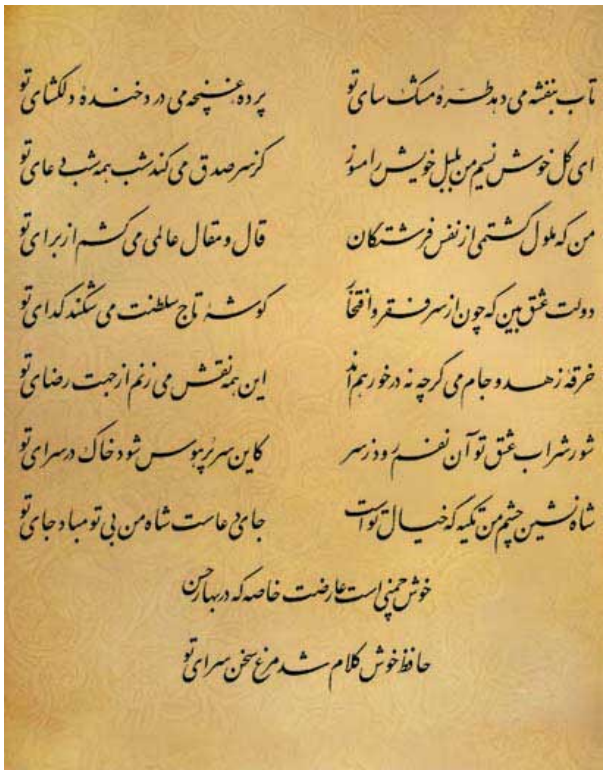


Figure 28: Ghazal from Class Mid-Age

Following are the cluster of most probable LDA terms with reference to Class b' as classified by both our two-phase and champion models:
çaʃm eye, *kaz* that from, *məy* wine, *gol* flower, *pardə* curtain and *del* heart.
We then looked at the distance networks that are ordered left to right, top to bottom in Class a', b' and c', to determine whether this poem's network of topic distance structure (Figure 29) most resembles the one our champion classifier predicted: Class 4 (Mid-Age).

Again, we found that the last image most resembles the second, in that there are two isolated topics that relate to each other!
Both our expert human judges agreed that this poem belongs to Houman's Class 3, Maturity. Classes 3 and 4 are chronologically adjacent and are both within class b'. This means that model prediction is close to our experts' choice.

6.3.6 Poem Example Four

Our champion classifier classified next poem shown in Figure 30 as Youth, while the two-phased model classified it as Maturity.

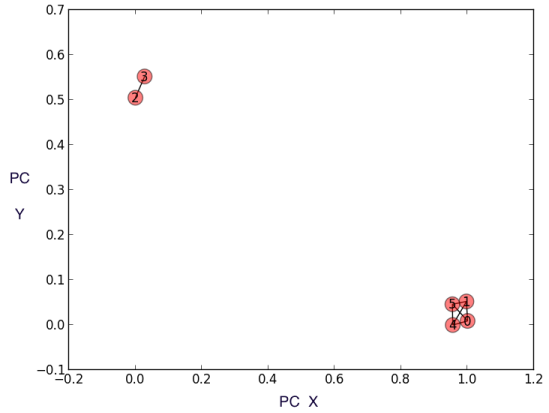
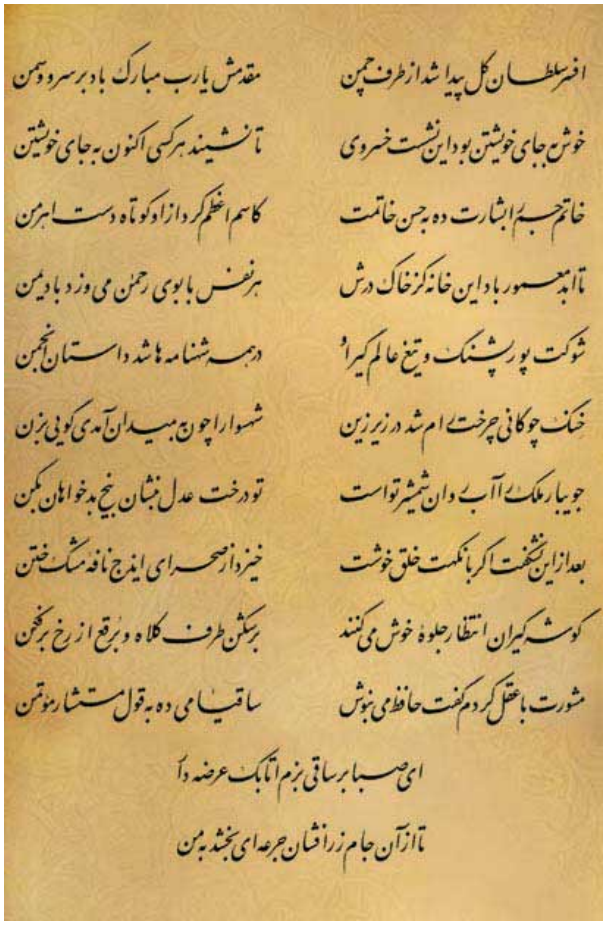


Figure 29: Poem Three's Topics Network

Officers of King of the flowers the grass adorn
The meadows welcome O God, the newly born.
What a pleasant gathering was this royal feast
Each one is seated upon his own throne.
Let your Seal, seal the fate of the Royal Seal
With your name, Satan's hands are cut and torn.
This house is eternally the gateway through which
The winds of compassion are fragrantly blown.
Glory of the Mighty King, his mythic sword
Book of Kings, and its readers have all sworn.
Tamed the stallion of fate, put under saddle
Mighty Rider played polo, the ball is thrown.
In this land flowing waters became your sword
Planted seeds of Justice, and evil intent forlorn.
No wonder, with your goodness, if from now on
From deserts, upon the breeze musk is flown.
Hermits patiently await your good vision
Raise your hat, throw aside the mask you've worn.
Sought counsel of my mind, said, Hafiz, drink!
Listen to my trusted friend, pour me wine until the morn.
Gentle breeze bestow this feast with plentiful horn
The bearer, with a cup or two, those like me may scorn.

Here are the cluster of most probable LDA terms with reference to Class a', as classified


افسر سلطان گل پیدا شد از طرف چمن
مقدمش یارب مبارک باد بر سرو و سمن

خوش جای خویشتن بود این نشست خسروی
تا نشیند هر کسی اکنون به جای خویشتن

خاتم جم بشارت ده به جم خاتمت
کاسم اعظم کرد از او کوتاه دست اهرمن

تا ابد معمور باد این خانه کز خاک درش
هر نفس با بوی رحمن می وزد باد یمن

شوکت پور پشنگ و تیغ عالم گیر او
در همه شنامه ها شد داستان انجمن

خنگ چوگانی چرخت ام شد در زیر زین
شهسوارا چون به میدان آمدی گویی بزن

جویبار ملک را آب و دان شمشیر تواست
تو درخت عدل بنشان بیخ بد خواهان بکن

بعد ازین نشگفت اگر با نکهت خلق خوشت
خیزد از صحرای ایذج نافه مشک ختن

گوشه گیران انتظار جلوه خوش می کنند
برشکن طرف کلاه و برقع از رخ برفکن

مشورت با عقل کردم گفت حافظ می بنوش
ساقیا می ده به قول مستشار مؤتمن

ای صبا بر ساقی بزم اتابک عرضه دار
تا ازآن جام زرافشان چرعه ای بخشد به من

Figure 30: Ghazal from Class Maturity

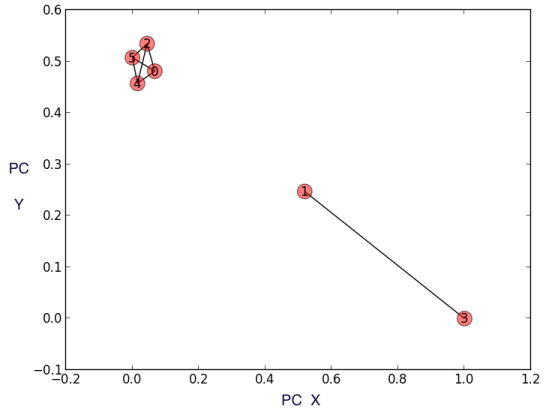


Figure 31: Poem Three's Topics Network

by both our two-phased and the champion models:
jam Jam, *zar* gold, *sâqi* maid, *jâm* and *məy* wine.
We next look at the distance networks that are ordered left to right, top to bottom in Class a', b' and c', to determine if this poem's network topic distance structure (bottom-right) most resembles with the one our champion classifier predicted: Class 1 (Youth):

We observe that the last image most resembles with the first, in that there are two further topics that only relate to each other.
Human expert 1 agreed with the machine, while expert 2 designated this poem as belonging to class 2, Post-Youth. Expert 2 presented evidence that there are two references to characters of Hafez's Post-Youth time in this poem. First, the Premier of Lorestan, Atabak Pashang (1355-1390), who lived during the period of Mobarezzedin, while the second, *ravâgə manzarə tʃaʃm* is a metaphor referring to Shah Shoja during the same period.
We asked the experts to discuss this. Expert 1 agreed with the presence of these references in the ghazal, but argued that "Pashang" refers to "poure Pashang", which is the mythological character *afrâsiâb* in Shahname by Ferdowsi. Expert 1 continued that Hafez could not have praised "Pashang ", using the double meaning the phrase implies, which both experts agreed on unless Hafez wrote this ghazal in the period before "Pashang" did harm by joining the Mobarezzedin, who blinded Atabak Nourelverd in lunar calendar 757. As we know, Hafez disliked Amir Mobarezzedin, and referred to him in his poetry as "Mohtaseb" or police. Hafez was close to Sheik Abu Eshegh, as was Atabak Nourelverd. The fact that Nourelverd was blinded by Atabak Pashang around 757 is evidence that the ghazal that praises Pashang must have been written before that horrible event took

place. It is highly unlikely that Hafez would have praised Atabak Pashang in this ghazal had he known of this. Hafez appreciated Abu Es-haagh bluntly:
râsti xâtam firuzəyə bu-eshâqi
xosh derakhshıd valı dolatə mostaaːjal bʊd.
In this poem, Hafez associates Abu Eshaagh with turquoise, which shines, but that his time was just too short! In this case, machine predicted this poem as Youth, expert 1 similarly voted for Youth and expert 2 voted for towards the end of class Youth. The experts' reasoning reaffirmed that our classifier's predictions were equal or very close to theirs even though the experts based their opinions on much more nuanced criteria than what an ML classifier could be capable of.

6.4 Clustering Semantic Analysis

Each poem's new label provided new perspective and insights, to enable us to interpret each Hafez's poem better, by associating it with the semantic characteristics of its associated cluster, in conjunction with its Houman classification. We could visualize the corresponding cluster, using *LDAvis* topic modeling (Sievert and Shirley, 2014a) who introduced and used the *Relevance* measure. Chuang et al. (2012) defined and developed *Saliency* as part of the Termite visualization tool.
For example, we selected to analyze a poem, number 230 from the Houman-labelled portion of the corpus, which was number 143 in Ganjour[54]. On the one hand, we saw that this poem belonged to class 5 or *before-senectitude* in Houman's classification. We looked at the top 30 terms of topic 3 which is at the center (darker coloured circle) in PCA depiction of 5 LDA topics, as we chose only 5 topics for better intuition purposes of perceived topics of a single poem in this LDA visualization. Figure 32 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 and relate to the poem.
Interacting with the visualization tool revealed other themes associated with this poem previously known as *before-senectitude*, 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

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

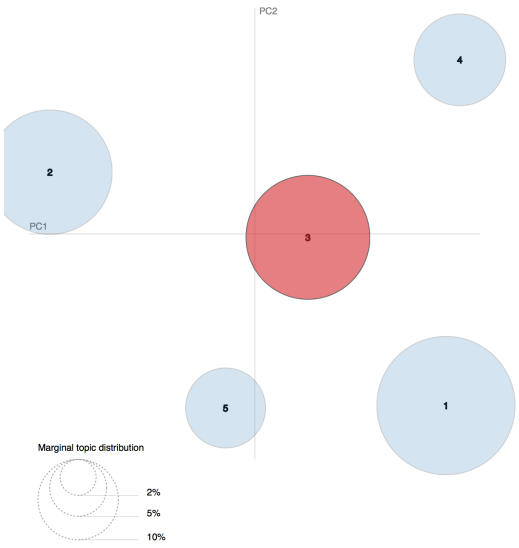


Figure 32: Intertopic Distance Map

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 used these words more metaphorically and mystically according to Houman.

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[55]
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.

[55]Mr. Shahriar Shahriari, our favourite translator, chose to implant a contemporary term.

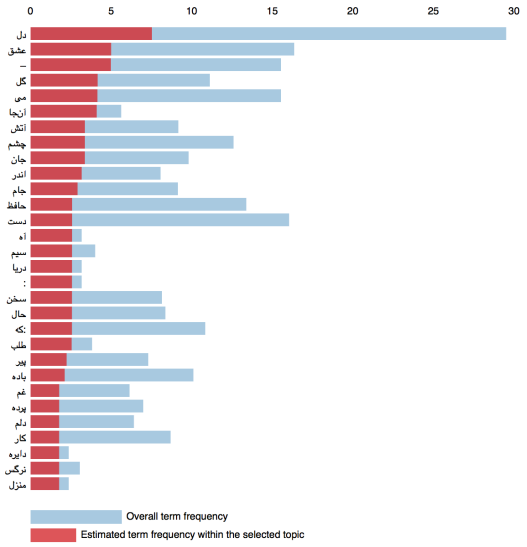


Figure 33: Top 30 Most Relevant Terms

6.4.1 Chronological Topic Terms Visualization

Sievert and Shirley (2014a) used LDA and introduced *relevance* of a term to a topic to develop an interactive visualization tool[56]. For brevity, we only mention LDA topics that were distinctively far from each other in the inter-topic distance map, Figure 34. Similarity, we calculated LDA topics for our corpus, for each class separately, in order to allow visualizations.

The class *Youth* has the following main topic terms:

- Topic 1: Dust [χâk], Circle [halqɛ], Lovers [ɔʃʃâq], Guru [piir], God [χodâ], Soul [jân], [bâdɛ] Wine;
- Topic 2: Blood [χun], Eyes [ʒolf], Joyous [χoʃ], Flaw [êib, Dream [χâb], ;
- Topic 3: Blood [χun], Promise [ahd], Articulation [soχən], Wineglass [jâm], Jewelry [sîm o ʒar];
- Topic 4: Wineglass [jâm], Assembly [majles], Deficiency [əib], Harp [tʃang], Angel [fereʃtə], Smooth-Tongued [ʃirin ʃoχan];
- Topic 5: Destiny [taqdîr], Laughter [χandə], Epistle [nâmə], Universe [jahan], Love [əʃq], Ocean [darîâ];

The class *Before Mid-age* has the following main topic terms:

[56]We used LDAvis library.

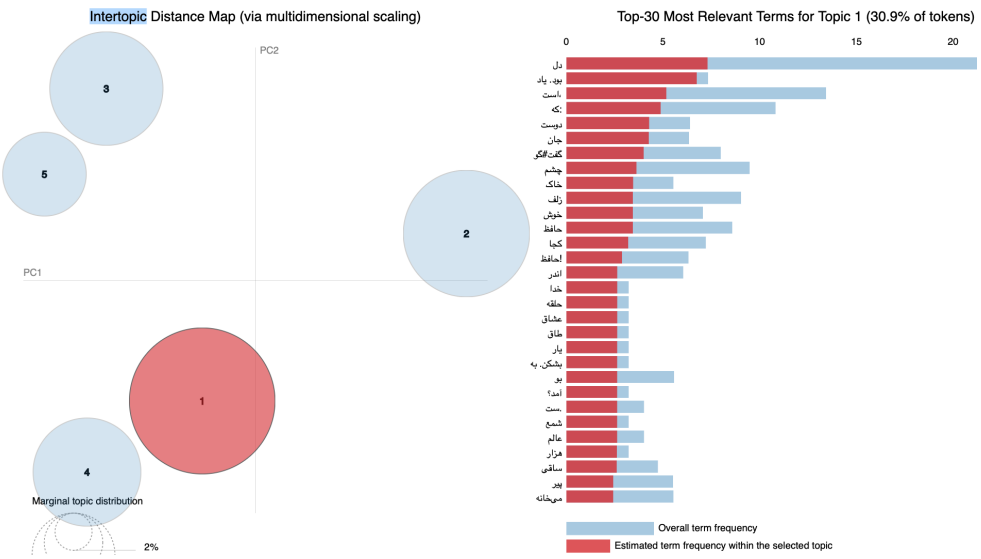


Figure 34: Houman Youth Class: Topic 1

- Topic 1: Universe [jahan], Dawn [sahar], Worth [qadr], Love [əʃq], Foundation [boni̊âd], Dilemma [gerəh], God [χodâ], Guru [pir], Hidden [nahân], Unreliable [sost], Psychologic [darun], Blood [χun];
- Topic 2: Cloak [χərqə], Sorrow [qam], Jewelry [zar o siim], Soul [jân], Fire [âtaʃ], Dust [χâk], Alley [kui̊], Blame [sarzaneʃ];
- Topic 3: Looking [nazar], Enjoyment [ei̊ʃ], Eyes [didə], Sign [neʃân], Wineglass [qadah], Epistle [nâmə], Maestro [motrəb];

The class *Mid-age* has the following main topic terms:

- Topic 1: Kandle [ʃamê], Joyous [χoʃ], Sorrow [qam], Looking [nazar], Ruin [χarabât], Night [ʃab];
- Topic 4: Governance [dôlat], Dust [χâk], Joyous [χoʃ], Satisfaction [kâm], Days [ai̊âm], Thought [andiʃə], Triumph [morâd], Eyes [tʃaʃm], Lover [âʃəq], Float [ravâ]n;
- Topic 5: Life [omr], Wineglass [bâdə], Attention [nazar], Joyous [χoʃ], Taste [tabĕ], Treasure [ganj];

The class *senectitude* has the following main topic terms:

- Topic 1: Old [pir], Love [əʃq], Sadness [qam], Dust [χâk], Eyes [tʃaʃm], Governance [dŏlat], Moghan [moqân], Ear [guʃ], Bird [morq], Desire [talab], Jewelry [lăl];

- Topic 3: Flower [gol], Love [əʃq], Universe [jahan], Sadness [qam], Saying [soχan], Look [nazar], Lover [âʃəq], Persistance [hemət];
- Topic 4: Desire [hâjat], Wineglass [jâm], God [χodâ], Love [əʃq], Cloak [χərqə], Moon [mâh], Image [naqʃ], End [âχər];

Certain terms are common among topics across all classes, such as: Hand [dast], Condition [kâr], Blood [χun], Heart [dəl], Wine [məy], Flower [gol] etc. In this section, we have shown a sample of automatically-generated topic terms associated with a few of the Houman classes. This is to show the extent of LDA model's ability to capture the high-level semantics of the Hafez poems. These terms are as close as we were able to get to Houman's perception of each class of Hafez poems, using our NLP techniques.

6.5 Houman vs. Raad Disagreements

In this section, we discuss the labels of two scholars on the same 249 poems of Hafez, using 4 chronological classes. Mr. Raad only perceived four classes, so we merged Houman's classes of 6 to 4 to be able to compare with Raad's. That is, to be able to compare the two scholars' labels, we decided to merge adjacent Houman labels $a + b$ and $e + f$, so that we could arrive at a logical set of four Houman classes. Kappa indicated a good level of agreement between the two scholars. The disagreements were in the poems that had elements of both classes. This claim was indirectly proven when the exclusion of disagreements improved the coherence. The other category of disagreements stemmed from the fact that Houman's perception of Hafez classes was more granular than that of Raad, therefore, the indicators that made Houman push certain poems to the extremes of Youth or Senectitude were not as strong or did not exist for Mr. Raad. It was easier for Mr. Raad to categorize such extremal poems as belonging to different classes.
If we categorize disagreements into two types of borderline and diverse, the analysis and comparison of the latter may sufficiently provide the gist of the difference between the two perspectives. In fact, the contrast between the two scholarly perspectives is insightful to help us understand Hafez's poetry better. For example, Houman ghazal number 38-169 (in Ghazvini) is the very last in group *Youth* but Raad has classified it as his group four or *Senectitude*.
I see no friends around whatever happened to every friend? I see no-one I love when did come to an end?
Water of life has turned dark where is glorious Elias? Flowers are all bleeding whence

the breeze which branches bend?
Nobody says that a friend has got the right to befriend What has come of loyalty? Whatever happened to every friend?
For years no gem has been dug from the mine of loyalty What happened to sunshine? And what about wind and rains trend?
This was the home of the Kings and the land of the Kind-hearted When did kindness end and since when Kings pretend?
The ball of compassion and joy is now inside the field Why is it that in this game still no players will attend?
Thousands of flowers are in full bloom yet not a song What happened to nightingales? Where did those thousands descend?
Venus is not making music any more did all her instruments burn? Nobody is in the mood to whom do the wine-sellers tend?
Hafiz secrets divine nobody knows stay silent Whom do you ask why isn't our turning fate now on the mend?

The collocation analysis of this Youth section of the corpus and selecting the top PMI measure in bigrams pick terms such as *joyous gem, friends come.* The elements of Houman's analysis in the ghazal are the earthy scene of desires, craving impulse, actual wine and flowers and birds, whereas for Raad the overall theme of the ghazal come across as depression, regrets comparing now with the good old days. The emotions that a contemporary Hafez scholar such as Raad would have perceived, was equivalent to what Houman had perceived of what an old Hafez might have felt. At the same time, this ghazal does not have the mysticism or introverted aspects that usually Houman expects to see Hafez picture in his later years.

Another example is the ghazal 50-151, labelled 2 (*before mid-age*) and 4 (*senectitude*) by Houman and Raad respectively.

A brief second spent grieving her loss is worth more than all the world.
Sell your sufi-robe for wine, it's good for nothing else.
The wine-dealers won't do business; whatever I have,
the prayer-mat of my stern devotions... isn't worth a cup.
The gate-keeper turned me away; what's happened
that I'm not worth the dust on your doorstep?
The sultan's crown holds the power of life and death—
it's attractive, sure, but not worth risking your head.

How easy it seemed at first, sailing the flood in quest of treasure...
Now, I wouldn't leave shore for a thousand pearls.
Better you should hide your face from your lovers,
the joy of victory's not worth the trouble of keeping captives.
In free surrender, struggle on like Hafiz, forgetting the world...
even if the least seed of effort should repay your weight in gold.

Again, degrading the prayer-mat and encouragement against sadness are the obvious patterns that Houman associated with the second period of Hafez's life in which he makes arguments against clergymen and their superficial religious campaigns. Raad, on the other hand, sees Hafez's disappointments, frustrations and demeaning comments towards a life that was not worth bonding with that only an old Hafez could have thought of while looking back. The collocation analysis of this section of the corpus and selecting the top PMI measure in bigrams pick terms such as *is sadness*. In the English translation above, the terms *risking* and *trouble* are used to reflect *worry* or *sadness* that can all be encapsulated in the term *ghəm* used in the original Persian.

6.5.1 Ontology Foundations of Hafez Ghazals

LDA-driven modelling defines relations among named entities. There are an average of 45 named entities per ghazal. Two terms are related if they belong to the same topic or the same class. However, two classes are also said to be *related* if they have common top terms. Top LDA terms are the ones with the highest membership probability within their topic. There are at least two categories of relations that are the potential cause of inconsistencies: direct and hierarchical relations. The *direct* relation is inter-class, regardless of the topic relations, meaning that two top terms are common between two classes. The *hierarchical* relation is intra-class topic-topic, meaning that a top term was not common directly, but was part of a topic that had common top-terms with a topic in other classes. Direct relation means that we go one level up in some tree and hierarchical relation means that we go two or more levels up. Our Hafez ontology is capable of querying both types of relations, but in this section, we only focus on visualizing *direct* relationships.

In this example, we show how the term ([xoʃ]) *Joyous* links or relates Youth Topic 2 to Mid-age Topic 1 directly, and how the term *Flaw* links or connects period Youth Topic 1 to period Mid-age Topic 4 and 5 indirectly or hierarchically through another term ([zolf] or [chaʃm]) meaning *eye*.

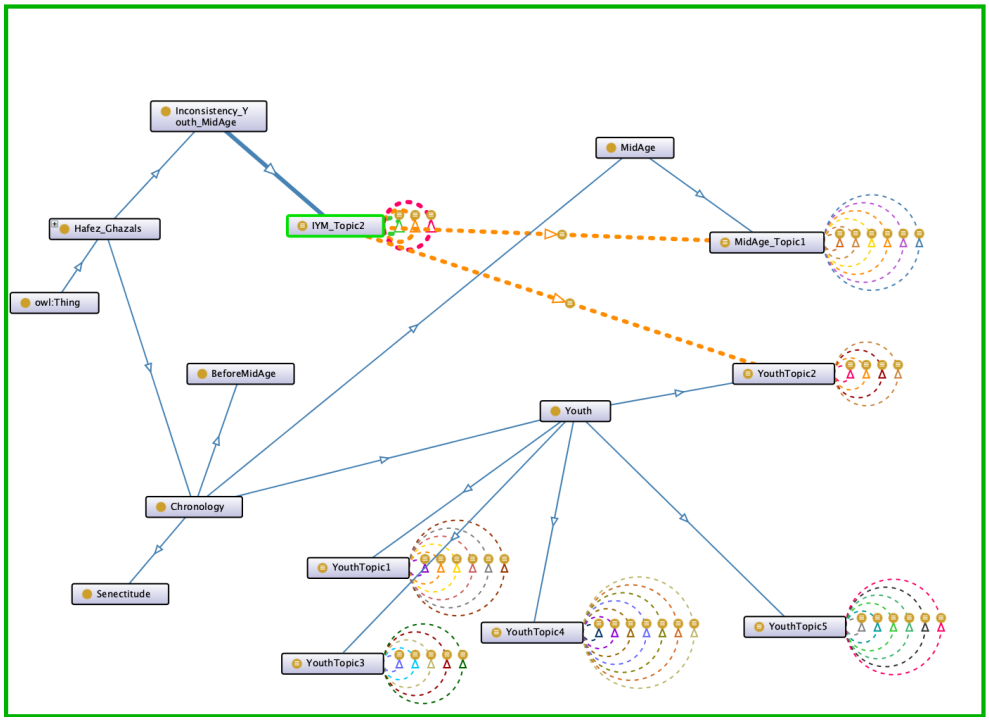


Figure 35: Inter-class Direct Relation: Joyous. Dotted lines show the common relation between two separate classes.

The DL Query (Description logic query is used in Protege (Musen, 2015)) to identify all relations associated with the entity (xoʃ) *Joyous* can be listed by running [*Joyous some*]. The coloured dotted lines depict the interrelations of classes through common topic terms. This visualization can provide clues to analyze inconsistencies among our Hafez experts and the classifier's predictions. We used Protégé for ontology development[57].

Another type of inconsistency is through intra-topic connections. For example, the word or entity əib *Flaw* is a top term in topic 1 in IYM_Topic[i] *Inconsistency*[58] that happens also to be part of topic 2 in class *Youth* that associates with another top term ʦhəshm *eye* that exists in multiple topics in class *Mid-age*. *Inconsistency* houses the LDA topic terms of the poems that Houman and Raad disagreed upon. The term *Flaw* is an indirect link that connects *inconsistency* class to both class *Youth* and *Mid-age*. In other words, there is a hierarchical relation that starts from a term in class *Inconsistency*, which coincides with another critical term *eye* in another class (*Youth*) that has other

[57]https://protege.stanford.edu/

[58]The IYM acronym stands for Inconsistency Youth Mid-age. DL query is showing the relationship terms common between Youth and Midage using "some", for topic *i* of Youth class.

common terms with different topics in a class *Mid-age*. For example, DL Query would be *IYM_ Topic2 EquivalentTo Eye some MidAge_ Topic1*[59]. The term Eye is common between the two classes of Youth and Mid-age. A *DL Query* result against Hafez ontology and its *OntoGraf*[60](Musen, 2015) visualization is shown in Figure 35. The dotted line depicts the common topic term as a relation which could be in turn, the source of inconsistency between the two classes.

Synonyms are an important attribute to consider in the current and future developments of a Hafez's ontology[61], and are mostly assumed out-of-scope in this thesis. The first version of an accessible Hafez ontology will be a future release. For example, the term nazar *Look* also belonged to *Inconsistency* class, topic 1, was semantically close to chaʃm *eyes* in meaning.

6.6 Conclusion

In this chapter, we used our visualization techniques using PCA and LDA. We used samples of poems from different chronological classes, and we looked at and compared their visual and semantic properties. We reviewed sets of topic terms associated with a particular class and poem. We studied the sample poems that were subject to the inter-annotator disagreement between our two experts, Houman and Raad, to help us shed some light on the obscure reasons behind their selections. We also tried to identify and visualize the topic terms and their relations that were associated with the choice of poems the scholars disagreed upon.

[59] According to Protégé documentation, the keyword 'some' is used to denote existential restrictions.
[60] Protégé plug-in.
[61] This ontology is a manual work-in-progress and its automation is a future work.

7

Conclusions and Future Work

The field of Digital Humanities, which bridges the gap between the humanities and computer science, includes automatic processing of text to facilitate literary research. From the perspective of NLP, our intention has been to achieve accurate classification for Hafez ghazals, which required preparing the corpus. We provided some information regarding the background of NLP that was relevant to poetry semantics. We also gained an understanding of the linguistic aspects of the Persian language and their impact on information retrieval. We conducted experiments to develop and adopt a reasonably performing classifier and visualized the topic models required for model checking. A summary of these achievements follows, in which we outline considerations for important future research opportunities to improve the model and apply it to other similar tasks.

7.1 Hafez Corpus

We made sure to properly clean and preprocess the data when preparing the corpus. The consistency and linguistic attention to detail that we devoted to our bilingual [62] corpora delivered classifications with satisfactory results, and continued refinements will improve

[62] Persian and English

the performance. Houman's methodology was our inspiration; hence, we followed his lead and the order of ghazals in our corpus for current and future use. We derived rules from Persian linguistics, then defined specifications and procedures and applied them to our Persian-English corpus during development.

7.1.1 LDA based *Similarity* features for SVM

We organized the main experiments into seven groups, explained each experiment and its associated evaluations and presented the important measures in tables. In the first experiment, we created the BOW training data and input it to the SVM classifier. For six classes, this experiment showed improvements over the baseline by about 5 points, to 37% accuracy. Adding the English section to the corpus increased the performance by another 2.5 points to almost 40% accuracy. Our observation of the confusion rate between the middle classes at that point indicated that combining pairs and lowering the number of classes might benefit the classification. The next group of experiments proved the hypothesis true, as the approach increased the accuracy to 61%.

In the third set of experiments, we used the Cosine similarity features to help the classifier better distinguish between classes b' and c', two of the amalgamated paired classes of the original six. This was due to our observation of the confusion matrix in the previous experiment, which indicated confusion between the two mentioned classes. We then added the similarity features, while retaining the BOW and LDA features as part of the SVM training data. This brought the accuracy of the classifier to almost 70%, and resolved the previous confusion.

In the fourth set of experiments, we decided to try the bilingual corpus for the three amalgamated classes, while keeping only the BOW feature. We then observed the effect of adding English to the previous Persian version, including the other features using the bilingual corpus, and compared the results. The bilingual method gave us 65% accuracy. Adding the LDA driven features to the bilingual corpus raised the accuracy to 73%. We noticed that the weakness in the confusion matrix was due to class a', so we applied the Cosine similarity features to that class and the accuracy jumped to 86%.

In an experiment, we decided to retain only the original c and d classes, as they accounted for the largest segment of corpus instances. The bilingual corpus with BOW gave us an improvement over the baseline to 68% accuracy. We also tried the LDA features standalone, which only improved over the baseline from 54% to 56%. However, keeping the BOW and adding the LDA driven features in this case drastically increased the classifier

accuracy.

At this point, we discussed whether the LSI or LDA-driven standalone similarity features would provide strong enough training features. Therefore, we created the training data with only normalized similarity features, once with LSI and once with LDA. The former resulted in 62% accuracy, while the latter reached 79%; to our surprise, this feature alone proved to be very powerful when training SVM classifiers. We reviewed our program design and methodology multiple times and verified that there was no overfitting or bias, by reviewing measures such as recall and precision. Since we had found this powerful method, in the final set of experiments (group seven) we returned to the original set of six classes and prepared the training data based only on the LDA-driven Cosine similarity features on all classes, and achieved an accuracy of almost 79%. We then applied the final model predictions for to visualization and analysis of the results and had some of the unlabeled ghazals validated by two experts.

7.2 Topic Visualization

We combined the six Houman classes into pairs for improving classification accuracy, built a cycle of top terms for each class and predicted the poem, and created graphical representations of each. We then compared the associated class graphs and terms with those for each predicted class of ghazals, to study the internal topic attributes. We used the Gensim Python library to analyze the results, and we hope this framework will help other researchers understand the Hafez poems.

We also applied the PCA method to the initialized LDA model, which allowed us to make 2D graphs for each class and ghazal. The graphs, in conjunction with the clusters of terms, not only inspired interesting discussions with the experts, but also played a critical role in the formation of ideas, and the rationale for predictions and their comparisons and interpretations.

7.3 Summary of Contributions

We created a reliable digital corpus of Hafez with proper and consistent linguistic properties, suitable for NLP activities. From Raad (2019), we collected secondary and more contemporary chronological labels, a new scholarly and labelling for Hafez ghazals and compared it with our pre-existing one in terms of consistency and semantic differences. We then proposed a classification refinement based on the inconsistencies between the

two labelling systems. Using one segment of disagreements, we brought the accuracy up by about 20%, while keeping the Doc2Vec feature set constant although the training set became slightly filtered and hence different. We precisely identified and excluded very few poems and this helped to drastically increase the accuracy. We used topic-terms *relevance* for visualization and analyzed the semantics of chronological classes. We also used LDA's topic-terms to establish the first Hafez ontology to further support scholarly analysis of Hafez poetry in conjunction with their twin chronological classifications. The contributions are as follows:

1. Hafez corpus development, linguistic refinements and preprocessing;
2. Semantic feature engineering;
3. The chronological classification of Hafez poetry (Rahgozar and Inkpen, 2016b);
4. The bilingual classification of Hafez poetry (Rahgozar and Inkpen, 2016a);
5. Homothetic clustering of Hafez poetry (Rahgozar and Inkpen, 2019);
6. Chronological semantics of Hafez poems (journal submission, Oxford DSH 2020):
    (a) Acquiring and semantic comparisons with a second Hafez scholar's chronological annotations;
    (b) Labelling inconsistency management, as a guide to improving classification of Hafez;
    (c) Using Doc2Vec (distributed memory) features in Hafez classifications;
    (d) LDA-driven ontology development of Hafez.
7. Poetry semantic visualization and tool development.

We developed a detailed architectural roadmap for the Poetry Information Extraction task, which essentially stems from semantic classification. Each classification component in the architecture had its evaluation methodology in place, following the ML best practices and evaluation standards. However, for the overall assessment, we also benefit from human judges and experts to approve or disapprove of the classifiers' predictions.

7.4 Future Work

There are many areas of improvement in the areas of semantic analysis, topic modelling, knowledge graph extraction and construction, linguistics and model visualization.
Our efforts can be the foundation for future research in the following fields:

1. Hafez semantic ontology refinements;
2. Semi-supervised classification of Hafez poetry;
3. Hafez poem by poem chronological sequencing and classifications using deep learning with a pre-trained BERT-style model for Persian;
4. Question/Answering support system for literary texts.

Though the LDA-driven Cosine similarity features provided us with an efficient means of semantic classification of poetic text, this framework requires a closer examination of its theoretical perspective. The mathematical interrelations of the LDA similarity with SVM, orchestrate well in our case. In addition, considering the random nature of LDA generation of the cycle of critical terms, it would likely be possible to develop a statistical procedure that can direct the experimental design for optimum classifications; perhaps it could help us arrive at more intuitive and distinctive topic terms. There may be room to improve on LDA models by more careful pre-processing of the corpus, to keep only highly relevant terms. Another research direction would be to develop LDA models that distinguish verbs from noun entities using POS features while extracting temporal chronological topics.
From the poetry viewpoint, particularly in the case of a high-calibre poet, we surmise that Hafez applied higher degrees of craftsmanship and poetic artistry during his late years. Such features, if captured, could correlate with chronological ordering to become another important aspect of his maturity and development process. In some of his works, Hafez's use of euphony is extremely strong. For example, when he pictures the fall, he skillfully chooses words that contain specific sounds without any concessions to the meaning. Hafez can make us picture the sound of wind carrying leaves over the ground with his meaningful, beautiful metaphors. Without invoking onomatopoeia, Hafez uses rhyme, alliteration, assonance, consonance and consonantal echo to encourage us to imagine a virtually real and deep-seated scene. It is as if one is watching a movie but on a much higher personal level. These clues in ghazals' feature engineering, again if captured, would help develop stronger classifications. For example, in the text, Hafez

beautifully interweaves the musical aspects of his poems, and given this, perhaps there are correlations to the maturity evolution of Hafez to musician-poet. Another important direction is gauging the effect of POS and entity-relation features in our 2-phase classification framework. This would provide a richer source of classification and insight in the chronological context.

Importantly, it seems the specific rhythm that Hafez has carefully developed is deliberately incorporated and intertwined with euphonic aspects of his ghazals. All these aspects are subsets of artistic layers, carefully built atop layers of metaphors that surround the meanings within messages. These concepts are essential ingredients in the makeup of Hafez's monumental poetry. Feature engineering of such attributes could reveal even greater import in conjunction with topic modelling and other Hafez classifications, if we could better link such deep properties to surface language.

For example, finding traces of prior poets' style and rhythms, given the fact that Hafez is the latest and the most preeminent in Persian poetry during about 500 years, along with Saadi Shirazi (13th century), Khakani (12th century), Dehlavi (13th and 14th century) and others. Ashoori (2009) strongly believes we can even find evident influences of important books such as *mersad-ol-ebad* (Lookout Servants) and *kaſſol-asrâr* (Secrets Revealed).

We believe that the pursuit and incorporation of these features in our ML process would provide different semantic perspectives. These features would provide diversified results with other multi-purpose semantic classifications, and create a much broader and deeper perspectives of Hafez's ghazal. Though the visualization of LDA and LDA-SVM classifications is a very new area, it has fruitful and exciting research potential. And, as we improve the filtering out of so-called *'irrelevant terms'* from LDA topic modelling, the visualization power will only increase. These two attributes will foster a combination of insightful presentation and characteristic analysis of topic model checking. Improved topic word clusters and more intuitive graphics will help human experts understand the reasoning behind improving ML automated topic models.

The semantic features that we extract could improve NLP classification tasks for many applications, such as authorship and authentication of concise texts. For example, one could distinguish the most authentic Hafez poems from the ones that are claimed to be his but are doubted by the scholars.

Persian Characters and Visualization Examples

The Variability of the Persian Alphabet

The location of a character in the word has implications for the word form. The flexibility in writing in Persian comes at a price of inconsistent word forms that pose a challenge in digitization and corpus development. Below we show the different notation in alphabetical forms of the same characters depending on their location in the word and the intricacies that could potentially result in inconsistent writing options in Figure 36.

Letter	UTF8 Code	Phonetics	Different images
ء	U+0621 Arabic letter Hamza		ء
آ	U+0622 Arabic letter Alef with Mad Combination of Hamza and Alef	ɑ:	آ - ـآ
ئ	U+0626 Arabic Hamza like Ye		ئ – ئـ - ـئ - ـئـ
ب	U+0628 Arabic letter Beh	b	ب – بـ - ـب - ـبـ
ة	U+0629 Arabic letter Teh Marbuta	t	ة
ت	U+062A Arabic letter Teh	t	ت – تـ - ـت - ـتـ
ث	U+062B Arabic letter Theh	s	ث – ثـ - ـث - ـثـ
ج	U+062C Arabic letter Jim	j	ج – جـ - ـج - ـجـ
ح	U+062D Arabic letter Hah	h	ح – حـ - ـح - ـحـ
خ	U+062E Arabic letter Khah	x	خ – خـ - ـخ - ـخـ
د	U+062F Arabic letter Dal	d	د - ـد
ذ	U+0630 Arabic letter Zal	z	ذ - ـذ
ر	U+0631 Arabic letter Reh	r	ر - ـر
ز	U+0632 Arabic letter Zeh	z	ز - ـز
س	U+0633 Arabic letter Seen	s	س – سـ - ـس - ـسـ
ش	U+0634 Arabic letter Sheen	ʃ	ش – شـ - ـش - ـشـ
ص	U+0635 Arabic letter sad	s	ص – صـ - ـص - ـصـ
ض	U+0636 Arabic letter Zad	z	ض – ضـ - ـض - ـضـ
ط	U+0637 Arabic letter Tah	t	ط – طـ - ـط - ـطـ
ظ	U+0638 Arabic letter Zah	z	ظ _ ظـ - ـظ - ـظـ
ع	U+0639 Arabic letter Ain		ع – عـ - ـع - ـعـ
غ	U+063A Arabic letter Ghain	q	غ – غـ - ـغ - ـغـ
ف	U+0641 Arabic letter Feh	f	ف – فـ - ـف - ـفـ
ق	U+0642 Arabic letter Qaf	q	ق – قـ - ـق _ ـقـ
ك	U+0643 Persian letter Kaf	k	ك – كـ - ـك - ـكـ
ل	U+0644 Arabic letter Lam	l	ل – لـ - ـل - ـلـ
م	U+0645 Arabic letter Mim	m	م – مـ - ـم - ـمـ
ن	U+0646 Arabic letter Noon	n	ن – نـ - ـن - ـنـ
ه	U+0647 Arabic letter Heh	h	ه – هـ - ـه - ـهـ
و	U+0648 Arabic letter Vav (WaW)	v as consonant	و - ـو
ي	U+064A Persian letter Yeh	J as consonant	ي – يـ - ـي - ـيـ
پ	U+067E Persian letter Peh	p	پ – پـ - ـپ - ـپـ
چ	U+0686 Persian letter Cheh	tʃ	چ – چـ - ـچ - ـچـ
ژ	U+0698 Persian letter zheh	ʒ	ژ - ـژ
گ	U+06AF Persian letter Gaf	g	گ – گـ - ـگ - ـگـ

Figure 36: Persian Characters

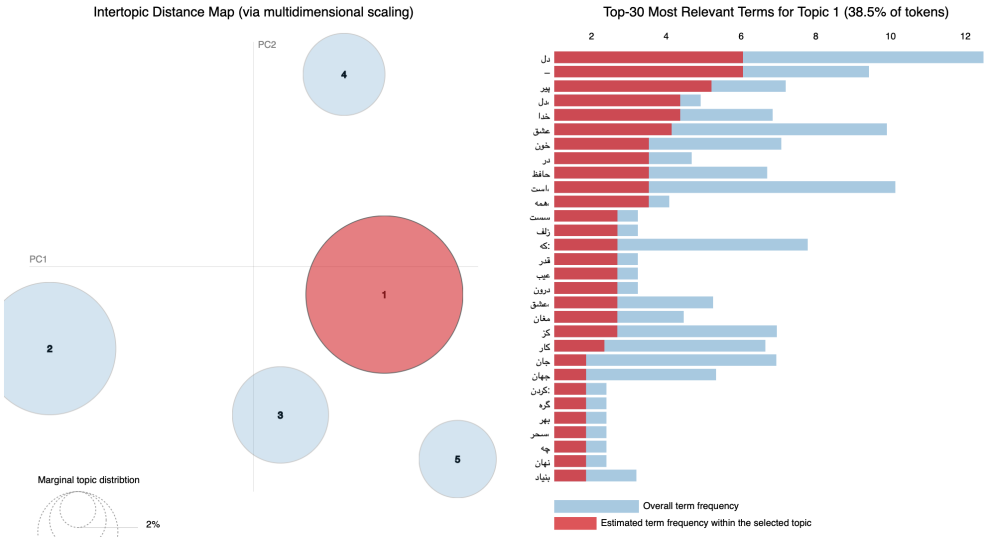


Figure 37: Houman Before Mid-Age Class: Topic 1

Visualization Examples

Figures 37, 38, 39 show the top terms associated with the topic belonging to different segments of the corpus. These visualizations are automatically generated by the LDA from the Persian corpus of Hafez. The English translations of the corpus are not complete.

Figure 38: Houman Mid-Age Class: Topic 1

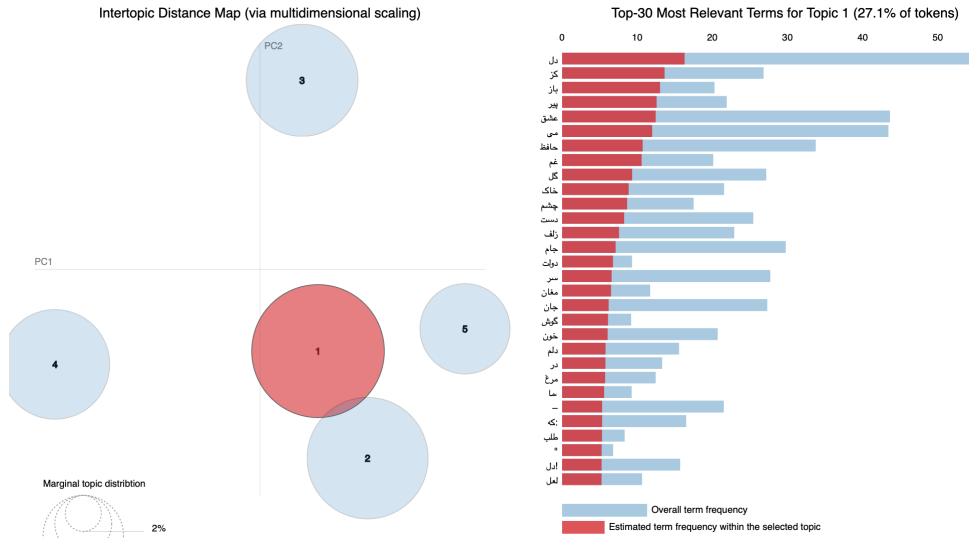


Figure 39: Houman Senectitude Class: Topic 1

Houman's, Raad's and Clustering Labels

The following three pages show Mr. Raad's labels against Houman's poems, indexed chronologically. The Houman index in column 1 correspond with his label counts in Table 2 in section 3.1.3. The Sim^2 column corresnponds with the champion homothetic clusters.

Houman	poem	Raad	Sim^2		Houman	poem	Raad	Sim^2
1	عشق‌بازي و جواني و شراب لعل‌فام،ا	1	4		125	محرمي حسبِ حالي ننوشتي، و شد ايامي چند.'	3	6
2	مدامم مست مي‌دارد نسيم جعد گيسويت؛	-	6		126	:ما شبي دست برآريم، و دعايي بكنيم	3	4
3	صبا، وقت سحر، بويي ز زلف يار مي‌آورد؛	1	4		127	،طاير دولت اگر باز گذاري بكند	3	6
4	كرشمه‌اي كن و بازار ساحري بشكن؛	3	2		128	صد ما دامن كشان هميشد در شرب زر كشيده،'	3	4
5	،ز اين خوش رقم كه بر گُل رخسار مي‌كشي	3	6		129	كه با وي گفت مسلمانان مرا وقتي دلي بود،'	4	3
6	بگرفت كار حسنت، چون عشق من، كمالي؛	3	1		130	،اگر نه باده غمِ دل ز يادِ ما ببرد	-	6
7	لبش مي‌بوسم و در مي‌كشم مِي؛	3	4		131	- به وقتِ گُل شدم از تو‌به‌ي شراب خجل	2	2
8	،دلم رميده شد، و غافل ام، منِ درويش	3	6		132	بيا و كِشتي ما در شطِ شراب انداز؛	4	2
9	به مژگان سيه كردي هزاران رخنه در دينم؛'	3	2		133	!اي كه در كوي خرابات مقامي داري	3	1
10	صنما! با غم عشق تو چه تدبير كنم؟	2	2		134	،گر بُوَد عمر، به مِي‌خانه رسم بارِ دگر'	3	1
11	ديده دريا كنم و صبر به صحرا فكنم؛	3	1		135	،بر سر آن ام كه گر ز دست برآيد	2	4
12	!اي خُرّم از فروغ رخت لاله‌زار عمر	3	4		136	خوش‌تر ز عيشِ صحبتِ باغ و بهار چيست؟	4	4
13	ه اي دل! گر از آن چاه زنخدان به در آيي،'	3	6		137	نفسِ بادِ صبا مشك‌فشان خواهد شد؛	3	1
14	:اگر به باده‌ي مشكين دلم كشد، شايد	3	1		138	رسيد مژده كه ايامِ غم نخواهد ماند؛	3	3
15	خوش‌تر از فكر مِي و جام چه خواهد بودن؟'	2	1		139	:مطلب طاعت، و پيمان، و صلاح از من مست	3	2
16	خيز تا از در مي‌خانه گشادي طلبيم؛	2	3		140	-صلاح كار كجا، و منِ خراب كجا؟	2	1
17	دُور ف صبح است، ساقيا! قدحي پر شراب كن:'	2	3		141	،هاتفي از گوشه‌ي مِي‌خانه دوش	3	2
18	-ز آن مِي عشق، كزاو پخته شود هر خامي'	3	1		142	:مطرب عشق عجب ساز و نوايي دارد	3	4
19	فتوي پير مغان دارم و قولي ست قديم	3	3		143	چو باد، عزم سر كوي يار خواهم كرد؛	4	1
20	ديگر ز شاخ سرو سهي، بلبل صبور	3	3		144	هواخواه تو ام – جانا! - و مي‌دانم كه مي'	3	5
21	صلاح از ما چه مي‌جوييِ؟ كه مستان را صلا'	3	1		145	!اي سروِ نازِ حُسن كه خوش مي‌روي به ناز'	3	1
22	بارها گفته ام، و بار دگر مي‌گويم	-	2		146	روي بنما، و وجودِ خويم از ياد ببر؛	3	2
23	روزه يك سو شد، و عيد آمد، و دل‌ها برخاس'	3	3		147	صبا! تو نكهت آن زلف مشك‌بو داري؛	3	1
24	رسيد مژده كه آمد بهار، و سبزه دميد؛	3	1		148	،در دير مغان آمد يارم، قدحي در دست	3	1
25	سَرم خوش است، و به بانگ بلند مي‌گويم'	3	1		149	"گفتم: "غم تو دارم"، گفتا: "غمت سر آيد'	3	2
26	كه عيب رندان مكن، اي زاهد پاكيزه سرشت!'	2	1		150	.دست در حلقه‌ي آن زلفِ دوتا نتوان كرد	3	6
27	گر مقامِ امن، و مِي بي غش، و رفيقِ شفيق'	4	6		151	عمري ست تا به راهِ غمت رو نهاده ايم؛	3	3
28	در ازل هر كاو به فيض دولت ارزاني بُوَد'	-	2		152	.روشن از پرتو رويت نظري نيست كه نيست	3	1
29	،در خرابات مغان گر گذر افتد بازم	1	2		153	،بوي خوش تو هر كه ز بادِ صبا شنيد	3	2
30	من ترك عشقِ شاهد، و ساغر نمي‌كنم؛	2	3		154	.اي پادشه خوبان! داد از غم تنهايي	3	1
31	:زاهد ظاهرپرست از حال ما آگاه نيست	1	3		155	چون خلوت‌گزيده را به تماشا چه حاجت است؟'	3	3
32	.ساقي! به نورِ باده برافروز جامِ ما	1	1		156	هر كه شد محرم دل، در حرم يار بماند؛	3	3
33	،آن غاليه خط گر سوي ما نامه نوشتي	3	4		157	.ز آن يار دلنوازم شُكري ست يا شكايت'	4	4
34	پيش از اينت بيش از اين انديشه‌ي عشاق ب'	1	2		158	ب باغبان گر پنج روزي صحبتِ گُل بايدش،'	4	6
35	ياد باد آن كه نهانت نظري با ما بود؛	1	1		159	راهي ست راه عشق كه هيچش كناره نيست؛	-	6
36	خوشا دلي كه مدام از پي نظر نرود؛	3	1		160	:فاش مي‌گويم و از گفته‌ي خود دلشاد ام'	4	1
37	،سلامي، چو بوي خوش آشنايي	2	1		161	اي صبا نكهتي از خاكِ ره يار بيار؛	3	2
38	ياري اندر كس نمي‌بينم؛ ياران را چه شد؟'	4	6		162	باز آي و دل تنگ مرا مونسِ جان باش؛	3	1
39	،اگر چه باده فرح‌بخش، و باد گل‌بيز است'	2	4		163	مژده، اي دل! كه دگر باد صبا باز آمد؛	3	2
40	- داني كه چنگ، و عود چه تقرير مي‌كنند؟'	2	3		164	، :حُسنت، به اتفاقِ ملاحت جهان گرفت	4	1
41	بُوَد آيا كه در مِي‌كده‌ها بگشايند؟	2	1		165	.در نظربازي ما بي‌خبران حيران اند	3	6
42	دوستان، دختر رز توبه ز مستوري كرد؛	2	1		166	.نه هر كه چهره برافروخت، دلبري داند	3	4
43	دوستان! وقت گُل آن به كه به عشرت كوشيم؛'	2	3		167	نيست در شهر نگاري كه دلِ ما ببرد؛	3	3
44	برو به كار خود اي واعظ! اين چه فرياد اس'	2	3		168	دل ما به دُورِ رويت، ز چمن فراغ دارد؛'	2	4
45	واعظان، كاين جلوه در محراب، و منبر مي'	2	6		169	دي، پيرِ مِي‌فروش - كه ذكرش به خير با'	2	4

3	4	باز آي، ساقيا! كه هواخواه خدمت ام؛	170		2	2	بيا كه قصر اَمَل سخت سست بنياد است؛	46
2	2	،ساقي ار باده از اين دست به جام اندازد'	171		3	-	تا راه اي بيخبر! بكوش كه صاحب خبر شوي.'	47
6	2	،به آب روشنِ مِي عارفي طهارت كرد	172		2	2	همي‌گفت اين م سحركه، رهروي در سرزميني،'	48
6	3	:عارف از پرتو مِي رازِ نهاني دانست	173		1	2	،غم زمانه كه هيچش كران نمي‌بينم	49
3	2	،عكس روي تو چو در آينه‌ي جام افتاد	174		2	4	دمي با غم به سر بردن، جهان يلنسر نمي‌ارز'	50
4	2	:با مدعي مگوئيد اسرارِ عشق و مستي	175		4	2	وقت را غنيمت دان، آن قَدَر كه بتواني:	51
1	2	،گر چه بر واعظ شهر اين سخن آسان نشود	176		3	2	:در خرابات مغان نورِ خدا مي‌بينم	52
4	2	- نقد ما را بُوَد آيا كه عياري گيرند؟	177		3	3	،گر مِي‌فروش حاجت رندان روا كند	53
6	2	ما بي‌غمانِ مستِ دل از دست داده ايم؛	178		2	2	برصبح است، و ژاله مي‌چكد از ابر بهمني؛'	54
2	2	:چو بشنوي سخنِ اهلِ دل، مگو كه خطا ست	179		4	4	- كه برد به نزد شاهان ز من گدا پيامي؟	55
2	2	ما نگوئيم بد، و ميل به ناحق نكنيم؛	180		1	3	اي دل! به كوي عشق گذاري نمي‌كني؛	56
5	4	بيا تا گُل برافشانيم، و مِي در ساغر اند'	181		2	2	.از ديده خونِ دل، همه، بر روي ما رود	57
3	2	بگذار تا ز شارع مِي‌خانه بگذريم؛	182		3	2	گر چه از آتشِ دل، چون خُمِ مِي، در جوش'	58
4	2	،به كوي مِي‌كده هر سالكي كه ره دانست	183		2	3	يوسفِ گم‌گشته باز آيد به كنعان، غم مخور'	59
3	2	تا ز مِي‌خانه و مِي نام و نشان خواهد بو'	184		6	2	،دلم رميده‌ي لولي‌وشي ست شورانگيز	60
3	3	صبا به تهنيتِ پيرِ مِي‌فروش آمد	185		3	3	روي بنما و مرا گو كه: "دل از جان بر گير'	61
4	-	شرابِ تلخ مي‌خواهم كه مردافكن بُوَد زور'	186		4	3	،من ام كه ديده به ديدار دوست كردم باز	62
3	3	به دُورِ لاله قدح گير، و بي ريا مي‌باش؛'	187		4	3	،هزار شكر كه ديدم به كام خويشت باز	63
4	3	ما درسِ سحر در رهِ مِي‌خانه نهاديم؛	188		4	3	.شراب و عيش نهان چيست؟ - كار بي بنياد'	64
6	2	:روزگاري شد كه در مِي‌خانه خدمت مي‌كنم'	189		1	3	ز آن رو المنّةلله كه در مِي‌كده باز است؛'	65
3	2	:حاشا كه من، به موسم گُل، تَركِ مِي كنم'	190		3	3	سحر ز هاتفِ غيبم رسيد مژده به گوش	66
4	2	!من و انكار شراب؟ - اين چه حكايت باشد؟'	191		1	3	،كنون كه بر كفِ گُل جام باده‌ي صاف است'	67
6	-	،گر من از سرزنش مدعيان انديشم	192		3	2	ساقيا! سايه‌ي ابر است، و بهار، و لب جوي'	68
6	2	:در همه دير مغان نيست چو من شيدايي	193		6	3	كنارِ آب، و پايِ بيد، و طبعِ شعر، و ياري'	69
3	2	من نه آن رند ام كه ترك شاهدِ و ساغر كنم'	194		2	3	مِي خواه و گُل افشان كن: از دهر چه مي"'	70
6	2	من ام كه شهره‌ي شهر ام به عشق ورزيدن؛	195		6	3	رونق عهد شباب است، دگر، بستان را؛	71
3	3	گر ز دستِ زلف مشكينت خطايي رفت، رفت؛	196		1	2	ساقي! بيار باده، كه ماه صيام رفت؛	72
6	2	،شراب بي غش، و ساقيِ خوش دو دامِ ره اند'	197		2	3	بيا كه، تُرك فلك خوان روزه غارت كرد؛	73
1	2	دل و دينم شد و دلبر به ملامت برخاست؛'	198		1	2	...دل مي‌رود ز دستم. صاحبدلان! خدا را'	74
3	3	اي بس نقد صوفي نه همه صافيِ بي غش باشد:'	199		3	2	من دوستدار روي خوش، و موي دلكش ام؛	75
2	3	صوفي ار باده به اندازه خورد، نوشش باد؛'	200		3	2	مرا مِهرِ سيه‌چشمان ز سر بيرون نخواهد ش'	76
4	3	صوفي نهاد دام و سر حقّه باز كرد؛	201		3	3	به كوي مِي‌كده، يا رب! سحر چه مشغله بود'	77
1	3	صوفي! گُلي بچين، و مرقع به خار بخش؛	202		6	3	يا رب! اين شمعِ دل‌افروز ز كاشانه‌ي كي'	78
2	3	صوفي! بيا كه خرقه‌ي سالوس بر كشيم؛	203		1	3	،آن سيه‌چرده، كه شيريني عالَم با او ست'	79
6	3	:صوفي! بيا كه آينه صافي ست جام را	204		2	4	اگر رَوَم ز پي‌اش، فتنه‌ها بر انگيزد:'	80
3	3	خيز تا خرقه‌ي صوفي به خرابات بريم؛	205		2	3	- ك شاهِ شمشادقدان، خسروِ شيرين‌دهنان،'	81
3	3	سال‌ها، دفتر ما در گروي صهبا بود؛	206		2	3	لعلِ سيرآب به خون تشنه، لب يارِ من است؛'	82
2	3	حافظِ خلوت‌نشين، دوش، به مِي‌خانه شد؛	207		2	3	كس نيست كه افتاده‌ي آن زلف دوتا نيست؛	83
4	3	من ام، كه گوشه‌ي مِي‌خانه خانقاه من است'	208		3	3	بعد از اين دست من، و دامنِ آن سروِ بلند'	84
3	-	اي دل! مباش يك دم خالي ز عشق و مستي؛	209		4	2	:دست از طلب ندارم، تا كامِ من بر آيد	85
4	2	كه سَرِ ارادتِ ما، و آستانِ حضرتِ دوست؛'	210		4	3	ز در در آ، و شبستان ما منور كن؛	86
1	3	ديده آئينه‌ا دل سراپرده‌ي محبتِ او ست.'	211		1	3	نس خيالِ روي تو در هر طريق همره ما ست:'	87
4	3	ساقي! بيا، كه يار ز رخ پرده برگرفت؛	212		6	3	اي شاهد قدسي! كه كشد بند نقابت؟	88
1	3	در نمازم خم ابروي تو با ياد آمد؛	213		1	3	صبا به لطف بگو آن غزال رعنا را	89
3	3	چو دست بر سر زلفش زنم، به تاب رود	214		6	2	.دردم از يار است و درمان نيز هم	90
1	3	رفتم به باغ، صبح‌دمي، تا چنم گُلي؛	215		6	1	:تو همچو صبح اي، و من شمع خلوت سحر ام'	91

2	3	بلبلي برگ گُلي خوش‌رنگ در منقار داشت؛	216		6	3	آن که از سنبل او غالیه تابي دارد	92
1	-	"صبح‌دم مرغ چمن با گُل نوخاسته گفت:	217		1	3	،آن که پامالِ جفا کرد، چو خاك راهم	93
3	3	،خستگان را چو طلب باشد و قوّت نبُوَد	218		3	3	.نفس بر آمد و کام از تو بر نمي‌آید	94
6	1	.دلا! بسوز، که سوز تو کارها بکند	219		1	3	،صبا اگر گذري افتدت به کشورِ دوست	95
3	-	ترسم که اشك، در غم ما، پرده‌در شود؛	220		4	3	،آن پیك نامور که رسید از دیارِ دوست	96
3	3	سحر با باد مي‌گفتم حدیث آرزومندي؛	221		2	-	،به جان او که گرم دست‌رس به جان بودي'	97
6	-	بنال، بلبل ! اگر با مَنَت سرِ یاري ست؛'	222		6	3	دلم جز مِهرِ مهرویان طریقي بر نمي‌گیرد'	98
2	2	:مزرع سبز فلك دیدم، و داس مه نو	223		1	3	دوش، در حلقه‌ي ما صحبتِ گیسوي تو بود؛	99
6	3	دیدي - اي دل! - که غم عشق دگر بار چه ك'	224		2	-	هرگزم ،نقش تو از لوحِ دلِ و جان نرود؛'	100
6	2	نو بهار است، در آن کوش که خوش‌دل باشي؛'	225		3	3	غلامِ نرگسِ مستِ تو تاج‌داران اند؛	101
2	3	،من کنون که مي‌دمد از بوستان نسیم بهشت،'	226		3	3	پیش تو گُل رو روشني طلعتِ تو ماه ندارد.'	102
2	4	یك‌کِي شعر تر انگیزد خاطر که حزین باشد؟'	227		3	3	گفت: "ب سحرم دولت بیدار به بالین آمد؛'	103
3	3	گرفتم باده، ب سحرگاهان، که مخمور شبانه،'	228		1	3	زلف آشفته، و خوي کرده، و خندان لب، و مس'	104
6	3	،بلبل، ز شاخ سرو، به گل‌بانگ پهلوي	229		6	2	گُل‌عذاري ز گُل‌ستانِ جهان ما را بس؛	105
6	3	سال‌ها، دل طلب جام جم از ما مي‌کرد؛	230		2	3	تا دوش مي‌آمد، و رخساره برافروخته بود؛'	106
6	-	به سِرِّ جامِ جم، آن گه، نظر تواني کرد'	231		3	4	،خدا، چو صورتِ ابروي دل‌کشاي تو بست	107
2	3	:در ازل، پرتو حسنت ز تجلي دم زد	232		1	3	نه خوش است خلوت، اگر یار یارِ من باشد؛'	108
6	3	حجاب چهره‌ي جان مي‌شود غبار تنم؛	233		1	3	خدا را، کم نشین با خرقه‌پوشان؛	109
1	1	.عشق تو نهالِ حیرت آمد	234		6	3	.زلف بر باد مده، تا ندهي بر بادم	110
3	3	.حاصل کارگهِ کون و مکان این همه نیست	235		3	3	اگر رفیق شفیق اي، درست پیمان باش؛	111
3	3	و ان دوش، وقت سحر، از غصه نجاتم دادند؛'	236		3	2	دارم از زلف سیاهش گِله چندان که مپرس؛'	112
4	2	"الا یا ایهاالساقي! ادر کاساً و ناولها"'	237		2	3	فکرِ بلبل همه آن است که گُل شد یارش؛'	113
6	2	،دوش دیدم که ملائك درِ مي‌خانه زدند	238		1	2	چه بودي، ار دلِ آن ماه مهربان بودي؟	114
6	4	که مرا به رندي، و عشق آن فضول عیب کند،'	239		4	3	،هزار جهد بکردم که یارِ من باش اي	115
1	4	پیرانه سرم عشق جواني به سر افتاد؛	240		1	1	و قتل این خسته به شمشیر تو تقدیر نبود؛'	116
2	4	هر هر چند پیر، و خسته دل، و ناتوان شدم،'	241		6	1	خود غلط بو ما ز یاران چشمِ یاري داشتیم؛'	117
1	4	.درخت دوستي بنشان: که کام دل به بار آرد'	242		6	3	ب به غیر آن که بشد دین و دانش از دستم،'	118
2	4	مژده‌ي وصل تو کو؟ - کز سر جان برخیزم؛'	243		3	3	ز دلي که غیب نماي است، و جامِ جم دارد،'	119
2	3	.معاشران! گره از زلف یار باز کنید	244		2	3	بنده شاهد آن نیست که مویي و میاني دارد:'	120
2	3	،آنان که خاك را به نظر کیمیا کنند	245		2	3	:شنیده ام سخني خوش که پیر کنعان گفت	121
1	3	.گداخت جان که شود کار دل تمام، و.. نشد'	246		3	3	،گر چه افتاد ز زلفش گره‌اي در کارم	122
1	4	،دو یار زیرك، و از باده‌ي کهن دو مني	247		1	-	آت سینه از آتش دل، در غم جانانه، بسوخت:'	123
3	4	حالیا مصلحت وقت در آن مي‌بینم	248		1	3	بي مِهر رخت روزِ مرا نور نماند است؛	124
6	4	اگر آن تُرکِ شیرازي به دست آرد دلِ ما ر'	249					

List of Definitions

ACE is a framework for knowledge representation, specification, and query language that can describe the vocabulary and the syntax, handle ambiguity and anaphoric references; also Automatic Content Extraction (ACE) is a research program for developing advanced Information extraction technologies.

CoreNLP Stanford CoreNLP integrates many of Stanford's NLP tools, including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the coreference resolution system, sentiment analysis, bootstrapped pattern learning, and the open information extraction tools[63].

DOLCE is a library of ontology foundations and stands for Descriptive Ontology for Linguistic and Cognitive Engineering, the main aim of DOLCE was to negotiate meaning of words for effective cooperation among multiple artificial agents.

Event Nugget (EN) A nugget is a predicated phrase on either an event or an entity. In other words, a nugget is a meaningful standalone text fragment.

Entities, Relations, and Events (Light/Rich ERE) is an annotation task that has evolved from lightweight treatment of entities, relations and events (ERE) to a richer representation of phenomena of interest (Song et al., 2015).

PropBank is both refers to the task of or to an actual corpus that is annotated according to verbal propositions and their arguments.

Protégé is an ontology editor and system by Stanford University.

Richer Event Descriptions (RED) is an extension to ERE for annotation.

VerbNet a digital dictionary or database that classifies verbs according to their semantics and syntactic behavior.

[63] https://stanfordnlp.github.io/CoreNLP/

References

Akiva, N. and M. Koppel (2013). *A generic unsupervised method for decomposing multi-author documents*, pp. 2256–2264. Journal of the American Society for Information Science and Technology, 64(11). (Cited on pages 24 and 27.)

Al-Shargabi, B., W. Al-Romimah, and F. Olayah (2011). A comparative study for Arabic text classification algorithms based on stop words elimination. In *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications*, pp. 1–5. (Cited on page 54.)

AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems 22*(5), 382–387. (Cited on pages 16 and 40.)

Aletras, N. and M. Stevenson (2013). *Evaluating topic coherence using distributional semantics*, pp. 13–22. Potsdam, Germany: Tenth International Workshop on Computational Semanctics. (Cited on page 26.)

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press. (Cited on page 9.)

Anuar, N. and A. B. M. Sultan (2010). Validate conference paper using DICE coefficient. *Computer and Information Science 3*(3), 139. (Cited on page 26.)

Arberry, A. J. (2004). *Fifty poems of Hafiz*. Routledge. (Cited on page 38.)

Ardanuy, M. C. and C. Sporleder (2015). Clustering of novels represented as social networks. *LiLT (Linguistic Issues in Language Technology) 12*, 1–31. (Cited on page 1.)

Artstein, R. and M. Poesio (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34*(4), 555–596. (Cited on page 79.)

Aryal, S., K. M. Ting, G. Haffari, and T. Washio (2014). MP-Dissimilarity: A data dependent dissimilarity measure. In *2014 IEEE International Conference on Data Mining*, pp. 707–712. IEEE. (Cited on page 24.)

Asgari, E. and J.-C. Chappelier (2013). Linguistic resources and topic models for the analysis of Persian poems. In *Proceedings of the workshop on computational linguistics for literature*, pp. 23–31. (Cited on page 79.)

Ashoori, D. (2009). *Mysticism and "Rendy"*. Markaz. (Cited on pages 34 and 130.)

Assi, S. M. (1997). Farsi linguistic database (FLDB). *International journal of Lexicography 10*(3), 5. (Cited on page 16.)

Baccianella, S., A. Esuli, and F. Sebastiani (2010). Selecting features for ordinal text classification. In *IIR*, pp. 13–14. (Cited on page 56.)

Basile, P., A. Caputo, and G. Semeraro (2009). UNIBA-SENSE@ CLEF 2009: Robust WSD task. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 150–157. Springer. (Cited on page 19.)

Baumann, S., T. Pohle, and V. Shankar (2004). *Towards a socio-cultural compatibility of MIR systems*, pp. 460–465. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 04). (Cited on page 21.)

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi (2011). Lessons from building a Persian written corpus: Peykare. *Language resources and evaluation 45*(2), 143–164. (Cited on pages 15, 27, and 40.)

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84. (Cited on pages 52 and 53.)

Blei, D. M. and J. Lafferty (2005). *Correlated topic models*, pp. 147–154. Vancouver, Canada: In Advances in Neural Information Processing Systems 17 (NIPS). (Cited on page 24.)

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022. (Cited on pages 23, 24, 25, and 52.)

Bouma, G. (2009). *Normalized (pointwise) mutual information in collocation extraction.*, pp. 31–40. Potsdam, Germany: In Proceedings of the Biennial GSCL Conference. (Cited on page 26.)

Brehmer, M., S. Ingram, J. Stray, and T. Munzner (2014). Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.

IEEE transactions on visualization and computer graphics 20(12), 2271–2280. (Cited on page 29.)

Brockhaus, H. (1875). *FA Brockhaus in Leipzig: Vollständiges verzeichniss der von der Firma FA Brockhaus in Leipzig seit ihrer Gründung durch Friedrich Arnold Brockhaus im Jahre 1805 bis zu dessen Hundertjährigem Geburtstage im Jahre 1872 verlegten Werke. In chronologischer Folge mit biographischen und literhistorischen Notizen*. FA Brockhaus. (Cited on page 39.)

Brown, P. F., S. A. DellaPietra, V. DellaPietra, and R. L. Mercer (1991). *WSD using statistical methods*, pp. 264–270. Berkeley, CA: Proceddings of the Annual Meetings of the Association for Computational Linguistics. (Cited on page 19.)

Carlo, C. M. (2004). Markov chain Monte Carlo and Gibbs sampling. *Lecture notes for EEB 581*, 251–259. (Cited on page 53.)

Celli, F., F. M. L. Di Lascio, M. Magnani, B. Pacelli, and L. Rossi (2010). Social network data and practices: the case of friendfeed. In *International Conference on Social Computing, Behavioral Modeling, and Prediction*, pp. 346–353. Springer. (Cited on page 24.)

Chaney, A. J.-B. and D. M. Blei (2012). Visualizing topic models. In *Sixth international AAAI conference on weblogs and social media*. (Cited on pages 28 and 29.)

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei (2009). *Reading tea leaves: How humans interpret topic models*, pp. 288–296. Vancouver, Canada: In Advances in Neural Information Processing Systems 21 (NIPS-09). (Cited on page 26.)

Choo, J., C. Lee, C. K. Reddy, and H. Park (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics 19*(12), 1992–2001. (Cited on page 29.)

Christensen, L. R., D. W. Jorgenson, and L. J. Lau (1975). Transcendental logarithmic utility functions. *The American Economic Review 65*(3), 367–383. (Cited on page 82.)

Chuang, J., C. D. Manning, and J. Heer (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of AVI'12, the ACM International Working Conference on Advanced Visual Interfaces*, pp. 74–77. (Cited on pages 28 and 116.)

Chuang, J., D. Ramage, C. D. Manning, and J. Heer (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 443–452. (Cited on page 28.)

Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational linguistics 16*(1), 22–29. (Cited on page 24.)

Colas, F. and P. Brazdil (2006). Comparison of SVM and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pp. 169–178. Springer. (Cited on page 45.)

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine learning 20*(3), 273–297. (Cited on pages 4 and 45.)

Cui, W., S. Liu, Z. Wu, and H. Wei (2014). How hierarchical topics evolve in large text corpora. *IEEE transactions on visualization and computer graphics 20*(12), 2281–2290. (Cited on page 29.)

Deerwester, S., G. Furnas, Dumais, and S. Landauer (1990). *Knowledge and natural language processing. (KBNL knowledge based natural language) (technical)*, pp. Vol.33(8), p.50(22). Programs with Common Sense. Commun. ACM. (Cited on page 51.)

Dehkhoda, A. A. (1994). *Dehkhoda encyclopedia*. Iran: Tehran University Publication. (Cited on pages 15 and 30.)

Dumais, S. T., T. A. Letsche, M. L. Littman, and T. K. Landauer (1997). Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, Volume 15, pp. 21. (Cited on page 16.)

Elworthy, D. (2001). Question answering using a large NLP system. In *AUTHOR Voorhees, Ellen M., Ed.; Harman, Donna K., Ed. TITLE The Text REtrieval Conference (TREC-9)(9th, Gaithersburg, Maryland, November 13-16, 2000). NIST Special Publication. INSTITUTION National Inst. of Standards and Technology, Gaithersburg, MD.; Advanced Research Projects Agency (DOD), Washington, DC.*, pp. 298. ERIC. (Cited on page 9.)

Erkinov, A. (2002). Manuscript collections of the former Uzbek Academy of Sciences Institute of Manuscripts (1978-1998). *Manuscripta orientalia 8*(1), 36–38. (Cited on page 38.)

Farkhund, I., H. Binsalleeh, B. Fung, and M. Debbabi (2010). *Mining writeprints from anonymous e-mails for forensic investigation*, pp. 56–64. Digital Investigation, 2010, Vol.7(1). (Cited on page 17.)

Fautsch, C. and J. Savoy (2010). Adapting the TFIDF vector-space model to domain specific information retrieval. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1708–1712. (Cited on page 11.)

Fell, M. and C. Sporleder (2014). Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 620–631. (Cited on page 22.)

Fleiss, J. L., J. Cohen, and B. S. Everitt (1969). Large sample standard errors of kappa and weighted kappa. *Psychological bulletin 72*(5), 323. (Cited on page 71.)

Forman, G. and I. Cohen (2004). *Learning from little: Comparison of classifiers given little training*, pp. 161–172. Springer Berlin Heidelberg: Knowledge Discovery in Databases: PKDD. (Cited on page 49.)

Frith, S. (1988). *Music for pleasure: essays in the sociology of pop*. Routledge New York. (Cited on page 21.)

Fuhr, N. and C. Buckley (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems (TOIS) 9*(3), 223–248. (Cited on page 11.)

Ghayoomi, M. (2012). Word clustering for Persian statistical parsing. In *International Conference on NLP*, pp. 126–137. Springer. (Cited on pages 14, 15, and 27.)

Ghayoomi, M. and S. Momtazi (2014). *Weakly supervised text categorization using topic modeling*, pp. 34–44. Tehran: Third Conference in Computational Linguistics. (Cited on pages 25, 26, and 53.)

Ghayoomi, M. and S. Müller (2011). Multi-token units and multi-unit tokens in developing an HPSG-based treebank for Persian. In *Proceedings of Fourth International Conference on Iranian Linguistics, ICIL4*. (Cited on page 15.)

Gieseke, F., A. Airola, T. Pahikkala, and O. Kramer (2012). Sparse quasi-Newton optimization for semi-supervised Support Vector Machines. In *ICPRAM (1)*, pp. 45–54. (Cited on page 79.)

Gliozzo, A. and C. Strapparava (2006). *Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization*, pp. 553–560. Sydney, Australia.: In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006). (Cited on page 17.)

Graham, N., G. Hirst, and B. Marthi (2005). Segmenting documents by stylistic character. *Natural Language Engineering 11*(4), 397–415. (Cited on page 27.)

Gretarsson, B., J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth (2012). Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST) 3*(2), 1–26. (Cited on page 29.)

Griffiths, T. L. and M. Steyvers (2004). *Finding scientific topics*, pp. 5228–5235. In Proceedings of the National Academy of Sciences, V101. (Cited on pages 24 and 53.)

Guo, Y. and M. Xiao (2012). Cross language text classification via subspace co-regularized multi-view learning. In *ICML'12: Proceedings of the 29th International Conference on Machine Learning*, pp. 915–922. (Cited on page 16.)

Haghighi, A. and L. Vanderwende (2009). *Exploring content models for multi-document summarization*, pp. 362–370. Boulder, USA: In proceeding of Human Language Technologies: Annual Conference of the Norht American Chapter of Association for Computational Liguistics (NAACL). (Cited on page 24.)

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18. (Cited on pages 46, 47, 51, and 58.)

Hasouri, A. (2005). *Another Glimpse ("Hafiz az nigahi digar")*. Uppsala. (Cited on page 33.)

Hastie, T. and R. Tibshirani (1998). Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.), *Advances in neural information processing systems*, Volume 10. MIT Press. (Cited on pages 49 and 51.)

Hayes, P. J. and S. P. Weinstein (1990). CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. In *IAAI*, Volume 90, pp. 49–64. (Cited on page 10.)

Hirjee, H. and D. G. Brown (2010). *Using automated rhyme detection to characterize rhyming style in Rap music*, pp. 121–145. Empirical Musicology Review 5(4). (Cited on page 21.)

Hoffman, M., F. R. Bach, and D. M. Blei (2010). Online learning for latent Dirichlet allocation. In *Advances in neural information processing systems*, pp. 856–864. (Cited on pages 72 and 81.)

Hofmann, T. (1999). *Probabilistic latent semantic indexing.*, pp. 50–57. Berkeley, USA: In Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. (Cited on page 24.)

Houman, M. (1938). *Hafez*. Tahuri. (Cited on pages ii, 2, 36, 38, 77, 84, 87, and 92.)

Hubert, L. J. (1978). A general formula for the variance of Cohen's weighted kappa. *Psychological Bulletin 85*(1), 183. (Cited on page 71.)

Hughes, J. M., N. J. Foti, D. C. Krakauer, and D. N. Rockmore (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences 109*(20), 7682–7686. (Cited on page 28.)

Hutchins, J. (2005). The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. *noviembre de 3265*, 102–114. (Cited on page 8.)

Inkpen, D. and A. H. Razavi (2014). Topic classification using latent Dirichlet allocation at multiple levels. *Int. J. Comput. Linguistics Appl. 5*(1), 43–55. (Cited on pages 5, 23, 24, and 46.)

Joachims, T. (1998a). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pp. 137–142. Springer. (Cited on pages 4, 12, and 45.)

Joachims, T. (1998b). Text categorization with support vector machines: learning with many relevant features. In *European Conference on Machine Learning (ECML)*, Berlin, pp. 137–142. Springer. (Cited on page 49.)

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 217–226. (Cited on page 26.)

Jolliffe, I. T. (2002). Choosing a subset of principal components or variables. *Principal component analysis SSS*, 111–149. (Cited on page 57.)

Kanerva, P. (1993). *Sparse distributed memory and related models*, pp. 50–76. MIT Press. (Cited on page 19.)

Kaufman, L. and P. J. Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*, Volume 344. John Wiley & Sons. (Cited on pages 79 and 81.)

Keerthi, S., S. Shevade, C. Bhattacharyya, and K. Murthy (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation 13*(3), 637–649. (Cited on page 51.)

Keogh, E., K. Chakrabarti, M. Pazzani, and S. Mehrotra (2001). Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pp. 151–162. (Cited on page 11.)

Khan, L., F. Iqbal, and M. Baig (2008). *Speaker verification from partially encrypted compressed speech for forensic investigation*, pp. 74–80. Digital Investigation, 2008, Vol.7(1). (Cited on page 17.)

Kim, S., H. Kim, T. Weninger, and J. Han (2010). *Authorship classification:A syntactic tree mining approach*, pp. 65–73. In Proceedings of the ACM SIGKDD, Workshop on Useful Patterns. (Cited on page 21.)

Koch, S., M. John, M. Wörner, A. Müller, and T. Ertl (2014). VarifocalReader—in-depth visual analysis of large text documents. *IEEE transactions on visualization and computer graphics 20*(12), 1723–1732. (Cited on page 29.)

Koppel, M., N. Akiva, and I. Dershowitz (2011). *Unsupervised decomposiiton of document into authorial components*, pp. 1356–1364. 49th Annual Meeting of the Association for Computational Linguistics: Human Language technologies (HLT) Vol.1. (Cited on page 27.)

Kowsari, K., K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown (2019). Text classification algorithms: A survey. *Information 10*(4), 150–218. (Cited on page 10.)

Kwok, J.-Y. (1999). Moderating the outputs of support vector machine classifiers. *IEEE Transactions on Neural Networks 10*(5), 1018–1031. (Cited on page 46.)

Landauer, K. Thomas, Dumais, and T. Susan (1997). *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge.*, pp. 211–40. Psychological Review, 1997, Vol.104(2), [Revue évaluée par les pairs]. (Cited on pages 19 and 55.)

Lau, J., D. Newman, S. Karimi, and T. Baldwin (2010). *Best topic word selection for topic labelling*, pp. 605–613. Beijing, China: In Proceedings of the 23rd International Conference on Computational Linguistics. (Cited on page 26.)

Lau, J. H., T. Baldwin, and D. Newman (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing (TSLP) 10*(3), 1–14. (Cited on page 26.)

Lau, J. H., P. Cook, D. McCarthy, D. Newman, and T. Baldwin (2012). *Word sense inductin for novel sense detection*, pp. 591–601. Avignon, France: In Proceedings of the 13th Conference of the EACL. (Cited on page 24.)

Lau, L. J. (1969). Duality and the structure of utility functions. *Journal of Economic Theory 1*(4), 374–396. (Cited on page 81.)

Le, Q. and T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. (Cited on pages 73 and 80.)

Lee, M. D., B. Pincombe, and M. Welsh (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the annual meeting of the cognitive science society*, Volume 27. (Cited on pages 73 and 80.)

Lesk, M. (1986). Lexical disambiguation using simulated annealing. In *Proceedings 1986 SIGDOC Conference*, pp. 24–26. (Cited on page 19.)

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark, pp. 37–50. (Cited on page 13.)

Li, H., A. C. Graesser, and Z. Cai (2014). Comparison of Google translation with human translation. In *The Twenty-Seventh International Flairs Conference*. (Cited on page 8.)

Luštrek, M. (2006). *Overview of automatic genre identification*, Volume Intelligent Systems, pp. E9. Jamova 39, Slovenia: Jozef Stefan Institute, Deparment of Intelligent Systems. (Cited on page 21.)

Manning, C. D. (1999). Foundations of statistical natural language processing/Christopher D., Hinrich Schultze. (Cited on page 9.)

Manning, C. D., C. D. Manning, and H. Schütze (1999). *Foundations of statistical natural language processing*. MIT press. (Cited on page 10.)

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*. Cambridge university press. (Cited on pages 10, 45, 49, and 74.)

Manning, C. D., M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60. (Cited on page 27.)

Marcialis, G. L. and F. Roli (2002). Fusion of LDA and PCA for face recognition. *Department of Electrical and Electronic Engineering, University of Cagliari, Piazza díArmi 8th Meeting of the Italian Association of Artificial Intelligence (AI*IA)*, 10–13. (Cited on page 5.)

Martínez, A. M. and A. C. Kak (2001). PCA versus LDA. *IEEE transactions on pattern analysis and machine intelligence 23*(2), 228–233. (Cited on page 5.)

Mayer, R., R. Neumayer, and A. Rauber (2008). Rhyme and style features for musical genre classification by song lyrics. In *Ismir*, pp. 337–342. (Cited on page 21.)

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. (Cited on page 25.)

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica 22*(3), 276–282. (Cited on page 71.)

Melišek, J. M. and M. O. Pavlovicová (2008). Support vector machines, PCA and LDA in face recognition. *J. Electr. Eng 59*(203-209), 1. (Cited on page 5.)

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint 3781*, 1301. (Cited on pages 73 and 81.)

Mikolov, T., A. Deoras, D. Povey, L. Burget, and J. Černockỳ (2011). Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 196–201. IEEE. (Cited on page 75.)

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119. (Cited on page 18.)

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM 38*(11), 39–41. (Cited on page 19.)

Mimno, D., H. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). *Optimizing semantic coherence in topic models*, pp. 262–272. Edinburgh, UK: Conference on Empirical Methods in Natural Language Processing (EMNLP). (Cited on pages 25 and 72.)

Minka, T. and J. Lafferty (2002). *Expectation-propagation for the generative aspect model*, pp. 352–359. CA, USA: In Proceedings of the 18th Conference on Uncertaintly in AI. (Cited on page 24.)

Mitchell, T. M. et al. (1997). Machine learning. (Cited on page 11.)

Montoyo, A., M. Palomar, G. Rigau, A. Suarez, and J. Artif (2005). *Word Sense Disambiguation System*, pp. 299. Intell. Res. 23. (Cited on page 19.)

Mosavi Miangah, T. (2009). Constructing a large-Scale English-Persian parallel corpus. *Meta: journal des traducteurs/Meta: Translators' Journal 54*(1), 181–188. (Cited on pages 16 and 20.)

Munková, D., M. Munk, and M. Vozár (2013). Influence of stop-words removal on sequence patterns identification within comparable corpora. In *International Conference on ICT Innovations*, pp. 67–76. Springer. (Cited on page 54.)

Musat, J., S. Velcin, and M. Rizoiu (2011). *Inproving topic evaluation using conceptual knowledge.*, pp. 1866–1871. IJCAI. (Cited on page 25.)

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI Matters 1*(4), 4–12. (Cited on pages 123 and 124.)

Nastase, V. and C. Strapparava (2013). Bridging languages through etymology: The case of cross language text categorization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 651–659. (Cited on pages 16 and 17.)

Navigli, R. (2009). *ACM Comput. Surv*, pp. 2. ACM 41. (Cited on page 19.)

Newman, D., J. Lau, K. Grieser, and T. Baldwin (2010). *Automatic evaluation of topic coherence*, pp. 100–108. LA, USA: Human Language Technologies: The 11th Annual Conference of the Norht American Chapter of the Association for Computational Linguistics (NAACL HLT). (Cited on pages 24, 25, 26, and 72.)

Ng, A. Y. and M. I. Jordan (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in neural information processing systems*, pp. 841–848. (Cited on page 49.)

Nichols, E., D. Morris, S. Basu, and C. Raphael (2009). Relationships between lyrics and melody in popular music. *ISMIR 10*, 471–476. (Cited on page 21.)

Noorinaeini, A. and M. R. Lehto (2006). Hybrid singular value decomposition; a model of human text classification. *International Journal of Human Factors Modelling and Simulation 1*(1), 95–118. (Cited on page 52.)

O'Connor, J. (2012). *NLP workbook: A practical guide to achieving the results you want.* Conari Press. (Cited on page 8.)

Pandian, A. and M. A. Karim (2014). *A study of Authorship Attribution in English and Tamil Emails*, pp. 203–211. India: Research Journal of Applied Sciences, Engineering and Techonlogy 8(2). (Cited on page 17.)

Pedersen, T. (2000). *A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation*, pp. 63–69. Seatle: Proceedings of the 1st Annual Meeting of the Norht American Chapter of the Association for Computation Linguistics. (Cited on page 19.)

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543. (Cited on page 18.)

Pilevar, M. T., H. Faili, and A. H. Pilevar (2011). Tep: Tehran English-Persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 68–79. Springer. (Cited on page 16.)

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola (Eds.), *Advances in kernel methods - support vector learning*. MIT Press. (Cited on pages 49 and 51.)

Pollard, C. and I. A. Sag (1994). *Head-driven phrase structure grammar*. University of Chicago Press. (Cited on page 27.)

Porteous, I., D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling (2008). Fast collapsed Gibbs sampling for Latent Dirichlet Allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 569–577. (Cited on page 24.)

Prettenhofer, P. and B. Stein (2010). *Crosslanguage text classification using structural correspondence learning.*, pp. 1118–1127. Uppsala, Sweden: In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010). (Cited on page 16.)

QasemiZadeh, B. and S. Rahimi (2006). *Persian in MULTEXT-East*, pp. 541–551. Iran: In Advances in natural language processing, 5th international conference on NLP, FinTAL. (Cited on page 16.)

Raad, M. (2019). Ziaratgah. Ottawa-Canada. (Cited on pages ii, 72, 91, 94, 95, and 127.)

Rahgozar, A. and D. Inkpen (2016a). Bilingual chronological classification of hafez's poems. In *Proceedings of the fifth workshop on computational linguistics for literature*, pp. 54–62. (Cited on pages 7 and 128.)

Rahgozar, A. and D. Inkpen (2016b). Poetry chronological classification: Hafez. In *Canadian conference on artificial intelligence*, pp. 131–136. Springer. (Cited on pages 7, 46, 90, and 128.)

Rahgozar, A. and D. Inkpen (2019). Semantics and homothetic clustering of Hafez poetry. In *Proceedings of the 3rd Joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, pp. 82–90. (Cited on pages 7 and 128.)

Rasooli, M. S., M. Kouhestani, and A. Moloodi (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 306–314. (Cited on page 15.)

Řehůřek, R. and P. Sojka (2010, May). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50. ELRA. http://is.muni.cz/publication/884893/en. (Cited on pages 29, 46, 54, 72, 73, 75, and 80.)

Rezapour, A., S. M. Fakhrahmad, M. H. Sadreddini, and M. Zolghadri Jahromi (2014). An accurate word sense disambiguation system based on weighted lexical features. *Literary and Linguistic Computing 29*(1), 74–88. (Cited on page 20.)

Rigutini, L., M. Maggini, and B. Liu (2005). *An EM based training algorithm for cross-language text categorization*, pp. 200–206. Compiegne, France: In Proceedings of the International Conference on Web Intelligence. (Cited on page 16.)

Röder, M., A. Both, and A. Hinneburg (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408. ACM. (Cited on page 72.)

Rosen-Zvi, M., C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS) 28*(1), 1–38. (Cited on page 27.)

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*, pp. 44. SICS dissertation series, 2006. (Cited on page 18.)

Sarrafzadeh, B., N. Yakovets, N. Cercone, and A. An (2011). Cross-lingual word sense disambiguation for languages with scarce resources. In *Canadian Conference on Artificial Intelligence*, pp. 347–358. Springer. (Cited on page 20.)

Schütze, H., D. A. Hull, and J. O. Pedersen (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 229–237. (Cited on page 21.)

Scott, S. and S. Matwin (1998). *Text classification using WordNet hypernyms*, pp. 45–51. In Proceedings of the Coling-ACL Workshop on Usage of WordNet in Natural Language Processing Systems. (Cited on page 21.)

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR) 34*(1), 1–47. (Cited on pages 9, 10, 11, and 12.)

Seraji, M. (2011). A statistical part-of-speech tagger for Persian. In *NODALIDA 2011, Riga, Latvia, May 11–13, 2011*, Volume 11, pp. 340–343. (Cited on page 16.)

Seraji, M., B. Megyesi, and J. Nivre (2012). A basic language resource kit for Persian. In *Eight International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May 2012, Istanbul, Turkey*, SwePub National Library of Sweden, pp. 2245–2252. European Language Resources Association: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). (Cited on pages 13, 14, 15, 40, and 41.)

Seyerlehner, K., M. Schedl, T. Pohle, and P. Knees (2010). Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, SSPK2. (Cited on page 22.)

Shi, L., R. Mihalcea, and M. Tian (2010). *Cross language text classification by model translation and semi-supervised learning*, pp. 1057–1067. Uppsala, Sweden: ACL 2010. (Cited on page 16.)

Shima, K., M. Todoriki, and A. Suzuki (2004). SVM-based feature selection of latent semantic features. *Pattern Recognition Letters 25*(9), 1051–1057. (Cited on pages 12 and 47.)

Sievert, C. and K. Shirley (2014a). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70. (Cited on pages 116 and 118.)

Sievert, C. and K. E. Shirley (2014b). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Volume 63-70. (Cited on page 29.)

Silva, T. C. and D. R. Amancio (2013). Discriminating word senses with tourist walks in complex networks. *The European Physical Journal B 86*(7), 297. (Cited on page 20.)

Simon, C. P. and L. Blume (1994). *Mathematics for economists*, Volume 7. Norton New York. (Cited on page 82.)

Simonton, D. K. (1990). *Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets*, pp. 251–264. Computers and the Humanities, 24(4). (Cited on page 21.)

Snyder, J., R. Knowles, M. Dredze, M. Gormley, and T. Wolfe (2013). Topic Models and Metadata for Visualizing Text Corpora. In A. for Computational Linguistics (Ed.), *Proceedings of the 2013 NAACL HLT Demonstration Session*, Volume 5-9. (Cited on page 28.)

Song, Z., A. Bies, S. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma (2015). From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pp. 89–98. (Cited on page 139.)

Steinley, D. (2004). Properties of the Hubert-Arable adjusted Rand index. *Psychological methods 9*(3), 386. (Cited on page 80.)

Stevenson, A. and C. Soanes (2003). *Oxford Dictionary of English*. Oxford University Press. (Cited on page 19.)

Stilo, D. L. (2005). Iranian as Buffer Zone Between the Universal Typologies of Turkic and Semitic. In E. A. Csató, B. Isaksson, and C. Jahani (Eds.), *Linguistic convergence and areal diffusion: case studies from Iranian, Semitic and Turkic*, pp. 35–63. Rutledge. (Cited on page 15.)

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*. (Cited on page 27.)

Viera, A. J., J. M. Garrett, et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med 37*(5), 360–363. (Cited on page 79.)

Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J. C. González-Cristóbal (2011). Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *Twenty-Fourth International FLAIRS Conference*. (Cited on page 45.)

Vinh, N. X., J. Epps, and J. Bailey (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research 11*(Oct), 2837–2854. (Cited on page 81.)

Wan, C., R. Pan, , and J. Li (2011). *Bi-weighting domain adaptation for cross-language text classification.*, pp. 1535–1540. Barcelona, Catalonia, Spain.: In Proceedings of the 22nd International Joint Conference on Artificial Intelligence. (Cited on pages 16 and 17.)

Wang, X., A. McCallum, and X. Wei (2007). *Topical n-grams: Phrase and topic discovery, with an application to information retrieval*, pp. 697–702. Omaha, USA: IEEE Internationa Conference on Data Mining (ICDM). (Cited on page 25.)

Widdows, D. and K. Ferraro (2008). Semantic vectors: a scalable open source package and online technology management application. In *LREC*, pp. 1183–1190. In Proceedings of the 6th Internatinal Conference on Language Resources and Evaluation. (Cited on page 19.)

Wiebe, J. M., R. F. Bruce, and T. P. O'Hara (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pp. 246–253. (Cited on page 73.)

Yarowsky, D. (1994). *Decision lists for lexical ambiguity resolution*, pp. 88–95. Las Cruces: Proceedings of the 32nd Annual Meeting of the Association for Computation Linguistics. (Cited on page 19.)

Zhang, J. and L. M. Yong Yan (1997). Face Recognition: Eigenface, Elastic Matching, and Neural Nets. In *Proceedings of IEEE*, Volume 85. (Cited on page 29.)

Zhang, X. and M. Lapata (2014). Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680. (Cited on page 75.)

Zhang, Y., R. Jin, and Z.-H. Zhou (2010). Understanding Bag-of-Words model: a statistical framework. *International Journal of Machine Learning and Cybernetics 1*(1-4), 43–52. (Cited on pages 11 and 16.)

Zhao, B. and E. Xing (2007). *Biligual topic exploration, word alignment, and translation*, pp. 1689–1696. Vancouver, Canada: In Advances in Neural Information Processing Systems(NIPS). (Cited on page 24.)