Multilabel Subject-based Classification of Poetry

Andrés Lou, Diana Inkpen and **Chris Tănăsescu (Margento)**
University of Ottawa
School of Electrical Engineering and Computer Science
800 King Edward, Ottawa, ON, Canada, K1N 6N5

Abstract

Oftentimes, the question "what is this poem about?" has no trivial answer, regardless of length, style, author, or context in which the poem is found. We propose a simple system of multi-label classification of poems based on their subjects following the categories and subcategories as laid out by the Poetry Foundation. We make use of a model that combines the methodologies of tf-idf and Latent Dirichlet Allocation for feature extraction, and a Support Vector Machine model for the classification task. We determine how likely it is for our models to correctly classify each poem they read into one or more main categories and subcategories. Our contribution is, thus, a new method to automatically classify poetry given a set and various subsets of categories.

Introduction

Poetry computational analysis is becoming more and more popular, though the field remains largely unexplored, as evidenced by the lack of a substantial body of work published (Kaplan and Blei 2007). Text classification methods, however efficient or at least effective when processing prose, often have to be modified and fine-tuned in a very different manner when dealing with poetry. While it may appear at first that any sort of in-depth analysis applied to poetry is a monumental task for a machine (because of the richness of meanings and information that can be contained in a single poem, a single verse, or sometimes even a single word), studies like those of Greene, Bodrumlu, and Knight (2010) and Kao and Jurafsky (2012) show that this is indeed possible, and that tasks such as machine translation and natural language generation can be carried out to a certain degree of effectiveness even when the data involved is poetry.

While poetry can be classified using many different evaluation metrics, such as subject, historical period, author, school, place of origin, etc, we focus entirely on a subject-based classification task, making exclusive use of the lexical content of each poem in our corpus to determine the categories to which it belongs.

Related Work

While there exists a volume of work related to computational poetry, the field is still relatively unexplored. Kaplan and Blei (2007) showed that it is possible to classify a group of poems in terms of style and to visualize them as clusters. Kao and Jurafsky (2012) showed that both concrete features, such as rhyme and alliteration, and abstract features, like positive emotions and psychological well-being, can be used to determine whether a certain poem was written in the style of prestigious, award-winning poets or amateur poets. Features such as rhyme or rhythm can be extracted and abstracted from verses into syllable-stress patterns for further processing, as shown by (2010) and (2010). Jamal, Mohd, and Noah (2012) used a Support Vector Machine model to classify traditional Malay poetry called pantun, which is a form of art used to express ideas, emotions and feelings in the form of rhyming lines. The authors classified the poems by theme; they also trained a classifier to distinguish poetry from non-poetry. A total of 1500 pantun divided into 10 themes and 214 Malaysian folklore documents were used as the training and testing datasets. This work is similar to our work since the themes are similar to our categories, but we also have subcategories, and our models use additional features.

We note that many of the resources employed in the crafting of poetry can indeed be processed, or "understood", by a machine, even if there are many gaps yet to be filled: Genzel, Uszkoreit, and Och (2010) point out that the task of preserving the form and meaning of a poem is an example of an area where machine translation might never replace a human translator, though they point out that there is work to be done in the field.

Classifying poetry

In this work, we focus on how the vocabulary of a poem determines its subject. While seemingly intuitive, this notion is a much more difficult task to perform than what it seems at first glance. As an example, let us consider the following excerpt from *The Love Song of J. Alfred Prufrock*, by T. S. Eliot:

> Let us go then, you and I,
> When the evening is spread out against the sky
> Like a patient etherized upon a table;
> Let us go, through certain half-deserted streets,
> The muttering retreats
> Of restless nights in one-night cheap hotels
> And sawdust restaurants with oyster-shells:

	No. of Poems	Fraction
Total	**7214**	%
Love	1250	17.3
Nature	2218	30.7
Social Commentaries	2258	31.3
Religion	848	11.8
Living	3103	43
Relationships	2524	35
Activities	1144	15.9
Arts & Sciences	1723	23.9
Mythology & Folklore	356	4.9

Table 1: The nine categories and the total number of poems in in our training set.

> Streets that follow like a tedious argument
> Of insidious intent
> To lead you to an overwhelming question ...
> Oh, do not ask, "What is it?"
> Let us go and make our visit.

As is the case with many modern and contemporary poems, the subject of this celebrated high modernist piece is problematic, elusive, and multilayered. The question of what category this poem belongs to has a nebulous answer. The title, while indicative, cannot be used to readily classify it as a "Love" poem. Furthermore, the fact that it belongs to a certain category such as "Love" does not imply that it does not belong to a different category as well, such as "Living", nor does it imply whether it belongs to a subcategory thereof, specifically, the subcategory of "Marriage & Companionship" (indeed, as we will see, unequivocal single categorization is rare). Furthermore, is the speaker's insistent urge to travel and discover (new?) places actually a facetious one, as some of his diction strongly suggests, and then what is the target of his irony? Are possibly capital existential questions as the one in the penultimate line muffled by the modern condition of pointless rambling, undiscriminating consumerism, and chronic disorientation? And where is the announced love in the "tedious argument" of the alienating placeless cityscape? The task of determining whether a poem belongs to any given number of categories and subcategories, by means of analyzing its lexical content, is the objective of our work.

Data

The Poetry Foundation's goal since its establishment in 2003 is "to discover and celebrate the best poetry and to place it before the largest possible audience."[1] The foundation is a large organization, and its website includes a corpus of several thousand poems categorized by subject, occasion, holiday, and several others.

The foundation is the successor to the Modern Poetry Association, founded in 1941 and the previous publisher of *Poetry* magazine. Today, the Poetry Foundation is one of the largest literary foundations in the world.

The corpus we used to train our classifying models was the Poetry Foundation's archive of poetry as of November 2014 [2]. We developed a method of parsing and downloading the poem embedded on the HTML of every page in the poetry archives. Thus we produced a large corpus of unprocessed documents (more than 7,000 poems), each one of them annotated with its author, title, and its subjects.

Tokenization is the process of breaking down a string of characters into substrings comprised of individual words and punctuation signs, called tokens. A token is a sequence of characters that we treat as a string; the vocabulary of a text is the set of tokens that appear in it. We do not focus on all tokens, but instead on **word types**, which are "the form or spelling of [a] word independently of its specific occurances in the text" (Bird, Klein, and Loper 2009).

As an example of a tokenization process, we consider the following verses of Edgar Allen Poe's *The Raven*:

> Once upon a midnight dreary, while I pondered, weak and weary,
> Over many a quaint and curious volume of forgotten lore–

Splitting these into tokens, we obtain the following set of unique types: ",", "–", "I", "Once", "Over", "a", "and", "curious", "dreary", "forgotten", "lore", "many", "midnight", "of", "pondered", "quaint", "upon", "volume", "weak", "weary", and "while".

Each poem in our corpus was tokenized and mined for types, a task from which we built a word list containing all the types in the corpus and the probability associated to each type. To reduce the dimensionality of our vector, we removed stopwords, punctuation signs, capitalization, and types that did not appear in the whole corpus at least twice. Thus, we were left with a word list containing 29,537 unique types.

Table 1 shows the total number of poems in our training set and the break-down of each category. Since a given poem may belong to more than one main category, the percentages do not add up to 100%.

Methodology

Our methodology involves three distinct phases: 1) Determining the number of categories and subcategories, and their nature, in which to place each poem; 2) Determine a method to extract relevant features from each document, and 3) Selecting an appropriate classifying algorithm.

Main Categories and Subcategories

The nine main categories as laid out by the Poetry Foundation's archive are as follows: "Love", "Nature", "Social Commentaries", "Religion", "Living", "Relationships", "Activities", "Arts & Sciences", and "Mythology & Folklore".

The same archive divides each main category into several subcategories, each of which do not appear outside their parent category. Because of time constraints, we only examine the subcategories of three main categories:

Love: "Desire", "Heartache & Loss", "Realistic & Complicated", "Romantic Love", "Classic Love", "Infatuation &

[1] http://www.poetryfoundation.org/foundation/about

[2] http://www.poetryfoundation.org/browse/

Crushes", "Unrequited Love", "Break-ups & Vexed Love", "First Love".

Living: "Birth & Birthdays", "Infancy", "Youth", "Coming of Age", "Marriage & Companionship", "Parenthood", "Separation & Divorce", "Midlife", "Growing Old", "Health & Illness", "Death", "Sorrow & Grieving", "Life Choices", "The Body", "The Mind", "Time & Brevity".

Mythology & Folklore: "Ghosts & the Supernatural", "Horror", "Heroes & Patriotism", "Greek & Roman Mythology", "Fairy-tales & Legends".

Feature Extraction

The content-based nature of the classification task makes it ideal to use two models to extract features from our corpus: Term Frequency-Inverse Document Frequency (tf-idf) as applied to a Bag-of-Words model, and Latent Dirichlet Allocation (LDA).

Bag of Word features Each word type is a feature for the classification and its value could be binary (1 if the word appears in the document and 0 if not) or based on frequency. We used **tf-idf**, because it was shown to work better in text classification tasks (Sebastiani 2002). tf-idf relates the frequency of a given word within a document to the frequency of the same word across all documents of a corpus, essentially determining how important the word is within the corpus. Several ways exist to calculate $\mathrm{tf}(t, d)$ of a given word; we used the simple approach of calculating the number of times the term *t* appears in a given poem *d*. idf is given by:

$$\mathrm{idf}(t, \mathcal{D}) = \ln \frac{N}{|\{d \in \mathcal{D} \mid t \in d\}|}$$

where N is the total number of documents in a corpus, d is a document belonging to the corpus set $\mathcal{D}$ and t is a term. Thus the set in the denominator represents all the documents in the corpus that contain the term t and the $||$ operator denotes the cardinality of the set. tf-idf is then given by:

$$\mathrm{tf} - \mathrm{idf}(t, d, \mathcal{D}) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, \mathcal{D})$$

LDA features Latent Dirichlet Allocation was first described by Blei, Ng, and Jordan (2003) as a "generative probabilistic model for collections of discrete data, such as text corpora" (Blei, Ng, and Jordan 2003). The idea of LDA is to represent each document in a corpus as a collection of topics, where each topic is characterized by a distribution over a set of words. LDA assumes the following generative process for each document $\mathbf{d}$ in a corpus $\mathcal{D}$ (Blei, Ng, and Jordan 2003):

- Choose an N number of words in the form a Poisson distribution.
- Choose $\theta \sim \mathrm{Dir}(\alpha)$
- For each word $\mathbf{w}_n$ in N:
  - Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$
  - Choose a word $\mathbf{w}_n$ from $p(w_n|z_n, \beta)$

With this and the diagram in Figure 1, we present the full probability equation as written by Savov (2009):

$$P(W, Z, \theta, \alpha, \beta) = \prod_{k=1}^{T} P(\varphi_k; \beta) \prod_{d=1}^{\mathcal{D}} P(\theta_d; \alpha) \prod_{w=1}^{N_d} P(Z_{d,w} \mid \theta_d) P(W_{d,w} \mid \varphi Z_{d,w})$$

where $P(Z_{j,w} \mid \theta_d)$ is the probability of picking a topic Z for a word w from a document d, given the topic proportion of d is θ_d, and $P(W_{d,w} \mid \varphi Z_{d,w})$ is the probability of picking word W for the w-th word in document d assuming we were drawing it from the topic distribution for topic $Z_{j,w}$.

In practice, the challenge of using LDA lies in either empirically or experimentally estimating the parameters from which the model would produce our corpus. For our task, we used the Gensim Python module to implement an LDA model using Gibbs sampling for parameter estimation. For a detailed analysis on utilizing this method, see (Griffiths and Steyvers 2004). LDA has been shown to be an efficient way of performing text classification tasks and has become a popular tool in different areas of the subject. See (Li et al. 2011) and (Zhou, Li, and Liu 2009) for examples.

The documents can be represented as a collection of topics and each word in each document is associated with a distribution of these topics. The topics look like clusters of words with certain probabilities /weights that reflect the importance of each word for the topic. Each document is assigned a number of topics, each having a certain probability/weight. Thus, the topics will be used as features for our classifiers, and each document will be represented by the topics assigned to it by LDA, while the values of the features are the assigned probabilities/weights.

Feature Selection We filtered the resulting feature set with a χ^2 ranking algorithm. **Pearson's χ^2 test** is a statistical test used to determine whether two events are independent of each other: the higher the χ^2 statistic, the more likely it is that the two events are dependent of each other. χ^2 is actually somewhat inaccurate when it comes to determining the level of independence between two events to one degree of independence, and it is prone to rank as dependent a number of features with little actual dependence; however, Manning, Raghavan, and Schtze (2008) showed that the noise produced by these is not important for a classification task as long as no statements about statistical dependence is made. Using this method, we kept the 1000 highest ranking word-types as features.

Classifiers

To build our classifiers, we used a Support Vector Machine model, namely Weka's SMO classifier with a polynomial kernel (Hall et al. 2009). A **Support Vector Machine** (SVM) is a model wherein input vectors are non-linearly mapped to a very high-dimension feature space, [w]here a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine (Cortes and Vapnik 1995). SVM has shown to be very efficient in text classification tasks, and has been a standard in the field for over a decade (Joachims 1998).

tf-idf	Accuracy	Precision	Recall	F-Measure	AUC
Love	0.888	0.883	0.888	0.873	0.711
Nature	0.831	0.83	0.831	0.823	0.764
Social Commentaries	0.809	0.806	0.809	0.798	0.738
Religion	0.908	0.896	0.908	0.895	0.678
Living	0.748	0.754	0.748	0.74	0.728
Relationships	0.769	0.775	0.769	0.749	0.697
Activities	0.875	0.864	0.875	0.85	0.642
Arts & Sciences	0.849	0.85	0.849	0.83	0.711
Mythology & Folklore	0.958	0.949	0.958	0.948	0.625
Average	**0.848**	**0.845**	**0.848**	**0.834**	**0.699**
Baseline	0.794	0.788	0.794	0.790	*0.616*

Table 2: Binary model output for each of the main categories using only bag-of-words. Baseline denotes the result obtained without feature selection. Note that using feature selection produces a considerably higher AUC value.

tf-idf + LDA$_{k=100}$	Accuracy	Precision	Recall	F-Measure	AUC
Love	0.888	0.884	0.888	0.873	0.71
Nature	0.831	0.829	0.831	0.822	0.764
Social Commentaries	0.807	0.804	0.807	0.797	0.737
Religion	0.909	0.898	0.909	0.897	0.682
Living	0.745	0.750	0.745	0.738	0.726
Relationships	0.770	0.777	0.770	0.750	0.699
Activities	0.875	0.865	0.875	0.851	0.644
Arts & Sciences	0.849	0.850	0.849	0.830	0.711
Mthology & Folklore	0.957	0.949	0.957	0.948	0.622
Average	**0.848**	**0.845**	**0.848**	**0.834**	**0.699**
tf-idf + LDA$_{k=500}$					
Love	0.887	0.884	0.887	0.872	0.709
Nature	0.832	0.83	0.832	0.823	0.765
Social Commentaries	0.806	0.802	0.806	0.795	0.734
Religion	0.91	0.899	0.91	0.897	0.678
Living	0.749	0.754	0.749	0.742	0.73
Relationships	0.770	0.776	0.770	0.751	0.700
Activities	0.874	0.865	0.874	0.849	0.638
Arts & Sciences	0.849	0.850	0.849	0.831	0.712
Mythology & Folklore	0.957	0.948	0.957	0.947	0.622
Average	**0.848**	**0.845**	**0.848**	**0.834**	**0.699**
tf-idf + LDA$_{k=1000}$					
Love	0.889	0.885	0.889	0.874	0.712
Nature	0.833	0.832	0.833	0.825	0.766
Social Commentaries	0.805	0.801	0.805	0.794	0.733
Religion	0.909	0.898	0.909	0.896	0.68
Living	0.751	0.757	0.751	0.744	0.732
Relationships	0.77	0.776	0.77	0.751	0.7
Activities	0.873	0.863	0.873	0.848	0.638
Arts & Sciences	0.851	0.852	0.851	0.833	0.715
Mythology & Folklore	0.958	0.949	0.958	0.948	0.628
Average	**0.849**	**0.846**	**0.849**	**0.835**	**0.700**

Table 3: Binary model outputs for each of the main categories using tf-idf and LDA.

The experiments we ran consisted of two separate tasks: the classification of poems into one or more of the nine main categories, and the classification of poems inside one or more subcategories belonging to a main category.

The binary nature of a SVM classifier meant that each document, given a category or subcategory a, had to be classified as either "belonging to a" or "not belonging to a". We therefore had to train several binary models, one for each category and each subcategory analyzed. Each model is evaluated using the standard measures: accuracy, precision, recall (all for positive values for each classifier), and area under the ROC curve (AUC)[3]. For our evaluation, we performed a 10-fold cross-validation (the data is split into $k = 10$ equal-sized subsets; 1 subset is used for validation

[3]The ROC curve plots the true positive rate against the false positive rate at various threshold settings.

Living tf-idf+LDA$_{k=500}$	Accuracy	Precision	Recall	F-Measure	AUC
Birth & Birthdays	0.976	0.975	0.976	0.97	0.648
Infancy	0.982	0.802	0.979	0.977	0.587
Youth	0.906	0.905	0.906	0.896	0.757
Coming of Age	0.951	0.953	0.951	0.935	0.611
Marriage & Companionship	0.955	0.957	0.955	0.941	0.614
Parenthood	0.924	0.928	0.924	0.908	0.678
Separation & Divorce	0.984	0.984	0.984	0.979	0.591
Midlife	0.973	0.940	0.973	0.973	0.516
Growing Old	0.902	0.909	0.902	0.875	0.618
Health & Illness	0.939	0.937	0.939	0.925	0.658
Death	0.859	0.864	0.859	0.847	0.766
Sorrow & Grieving	0.901	0.909	0.901	0.875	0.632
Life Choices	0.952	0.939	0.952	0.937	0.560
The Body	0.939	0.923	0.939	0.912	0.514
The Mind	0.929	0.929	0.929	0.905	0.570
Time & Brevity	0.882	0.885	0.882	0.868	0.750
Average	**0.935**	**0.921**	**0.934**	**0.920**	**0.629**

Mythology & Folklore tf-idf+LDA$_{k=100}$	Accuracy	Precision	Recall	F-Measure	AUC
Ghosts & the Supernatural	0.905	0.916	0.905	0.897	0.818
Horror	0.952	0.907	0.952	0.929	0.500
Heroes & Patriotism	0.810	0.851	0.810	0.779	0.692
Greek & Roman Mythology	0.960	0.676	0.69	0.673	0.626
Fairy-tales & Legends	1.000	1.000	1.000	1.000	0.000
Average	**0.925**	**0.870**	**0.871**	**0.855**	**0.527**

Love tf-idf+LDA$_{k=500}$	Accuracy	Precision	Recall	F-Measure	AUC
Desire	0.837	0.841	0.837	0.822	0.741
Heartache & Loss	0.892	0.901	0.892	0.861	0.616
Realistic & Complicated	0.816	0.822	0.816	0.798	0.720
Romantic Love	0.837	0.835	0.837	0.824	0.735
Classic Love	0.942	0.938	0.942	0.93	0.664
Infatuation & Crushes	0.884	0.882	0.884	0.873	0.751
Unrequited Love	0.915	0.913	0.915	0.893	0.615
Break-ups & Vexed Love	1.000	1.000	1.000	1.000	0.000
First Love	0.971	0.969	0.971	0.962	0.593
Average	**0.899**	**0.900**	**0.899**	**0.885**	**0.604**

Table 4: Binary model outputs for each of the subcategories of Living, Mythology & Folklore and "Love".

while the remaining $k-1$ are used as training data; the process is repeated k times, each time a different subset being used for training). Our results are shown in Tables 2-4.

Results and Discussion

The issue of data imbalance could not be sorted out without decreasing the size of our corpus; there is a disproportionally larger amount of instances under the "Living" category and a disproportionally smaller amount of instances under "Mythology & Folklore". Overall, the results are acceptable, with all AUC measures well above 0.6 but none over 0.8. Further repetitions of the experiments and fine-tuning the parameters of the SVM classifier do not significantly improve the data. The subcategories show overall similar results, while presenting scarcity as an additional limiting factor.

Main Categories

We performed an experimental run with the entirety of the word-types extracted from the corpus, without including the LDA models in our training data. Results are shown in Table 2. The average AUC of this run is the lowest of all our experiments with the main categories.

After performing feature selection, we performed two sets of experimental runs: using only Bag-of-Words to extract features, and integrating both Bag-of-Words and LDA. Our purpose was to determine the impact the latter would have on our results, since the literature has shown the model to be popular in classification tasks. Our results, however, show that, while tf-idf alone delivers better results for some categories and tf-idf+LDA delivers better results for others, the average AUC is identical between the models, with all other statistics leaning, however slightly, towards tf-idf+LDA. The results of tf-idf are shown in Table 2.

The Gibbs sampling method of estimating LDA parameters leaves the task of selecting the number of LDA topics, k, to the experimenter. We made three experimental runs of tf-idf+LDA and $k = 100, 500, 1000$. Results are shown in Table 3. We also attempted to fully represent our corpus as an LDA distribution of topics by using nothing but the

$k = 500$ number of topics in our feature-space; they clearly show that stand-alone LDA topics are insufficient for any useful practical result[4].

Subcategories

The three main categories we selected to perform an experimental run were "Living", "Mythology & Folklore", and "Love", the first being the largest category in terms of number of poems, the second being the smallest, and the third falling somewhere in between.

Table 4 present our results for the subcategories. The average AUC measurement for the subcategories is noticeably lower when compared to the main categories. This decrement reflects the relative scarcity of each subcategory, as there are much fewer instances with which to train each classifier. "Living" has the highest average AUC, which, again, reflects the relative scarcity of data for the subcategories of Love and Mythology & Folklore. The results suggest that a major increase in the available instances of each category or subcategory would further improve the performance of the classifier.

Conclusion and Future Work

We have shown a simple method of determining whether a given poem belongs to an established category by listing its vocabulary in relation to the frequency of each term that belongs to it. While the idea of poems that do not approach a single given topic is not controversial, the categories themselves are not a universal convention. The very existence (or lack) of content-based categories of any sort might sometimes be a point of contention against subject-based classification tasks. SVM classifiers with feature selection achieved the best results for the main categories and subcategories.

Future work focusing on the content of poems for classifying purposes should refine models and account for both poetic diction and form. Style and the use of metaphors are both content-based concepts that should also be used in the task of classifying poetry. Advances in metaphor comprehension and development, as shown by Levy and Markovitch (2012), show that metaphors represented as mapping functions over a feature-space are a viable tool to make a machine "understand" a concept. Advances in rhyme and rhythm analysis (Genzel, Uszkoreit, and Och 2010) – which we shall complement with our own work on both meter and more euphonic techniques (alliteration, assonance, slant rhymes, etc.) as well as established poetic forms (sonnets, villanelles, terza rimas, etc.)– are steadily paving the road for automatic classification in such a deeply human field as poetry.

> Never say, your lexicon exhausted,
> that for lack of content the lyre is silent
> There may not be a poet but
> There always shall be poetry
> —- Gustavo Adolfo Bécquer

[4] An average AUC of 0.508

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning*.

Cortes, C., and Vapnik, V. 1995. Support-Vector Network. *Machine Learning, 20, 273–297*.

Genzel, D.; Uszkoreit, J.; and Och, F. 2010. Poetic statistical machine translation: Rhyme and meter. In *Conference on Empirical Methods in Natural Language Processing*, 158–166.

Greene, E.; Bodrumlu, T.; and Knight, K. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Conference on Empirical Methods in Natural Language Processing*, 524–533.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Aacademy of Sciences of the United States of America*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA data mining software: An update. In *SIGKDD Explorations*, volume 11.

Jamal, N.; Mohd, M.; and Noah, S. A. 2012. Poetry classification using Support Vector Machines. *Journal of Computer Science* 8, Issue 9:1441–1446.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98*, volume 1398 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 137–142.

Kao, J., and Jurafsky, D. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Workshop on Computational Linguistics for Literature*, 8–17.

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

Levy, O., and Markovitch, S. 2012. Teaching machines to learn by metaphors. *Proceddings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Li, K.; Xie, J.; Sun, X.; Ma, Y.; and Bai, H. 2011. Multi-class text categorization based on LDA and SVM. *Procedia Engineering* 1963–1967.

Manning, C. D.; Raghavan, P.; and Schtze, H. 2008. *Assessing χ^2 as a feature selection method*. Cambridge University Press.

Savov, I. 2009. Latent Dirichlet Allocation for scientific topic extraction. Unpublished.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Zhou, S.; Li, K.; and Liu, Y. 2009. Text categorization based on topic model. *Atlantis Press* 398–409.