UNSUPERVISED MACHINE LEARNING IN 20TH CENTURY ATTRIBUTION OF ITALIAN POETRY: MONTALE OR CIMA?

Abdulaziz Khader
MehtA+Tutoring

Nimish Baweja
MehtA+Tutoring

Kabir Goel
MehtA+Tutoring

August 4, 2021

ABSTRACT

In this paper, authorship attribution in Italian poetry will be conducted to determine the authorship of a specific poem after poems with recognized authorship have been collected. When a poet writes, s/he creates a unique "signature" based on the words and sentences they chose to include. The model will attempt to identify that signature and attribute it to the correct author. The model in question is the K-Means model, which suits this project perfectly. In the experiment, a set of 115 Montale poems and 140 Cima poems were used. The highest accuracy value for the model is 85%.

1 Introduction

Authorship Attribution is a complicated task. These complications escalate to a higher level when dealing with poetry. It lacks a lot of the patterns that regular text follows. Extracting features from poems is more complicated because factors like sentence length, theme, and term usage changes with each poem. This project involved the analysis of Italian poetry, specifically from Eugenio Montale and Annalisa Cima. Montale was an influential 20th-century Italian poet. Annalisa Cima was his friend and fellow poet, whom he often used to send poems to. When Montale passed, Cima published the "Posthumous Diary", which she claimed was a compilation of poems Montale wrote before his death. A few years later, some scholars figured out that some of the poems were written by Cima, not Montale. Some critics claim that Cima actually wrote most of these poems. The task in this project was to see how much of the Posthumous Diary was actually written by Montale, and which poems were written by Cima.

2 Related Work

Within the topic of authorship attribution in poetry, there have been several other studies that have similarities and relevance to this one. The most significant study that was used for comparison is the one from the University of Bologna. This study involves the same exact topic and had the same exact goals for the study.This study was conducted 10 years ago, and the professor who conducted this study wanted a revision. Other research projects that were relevant to this one are the Arabic Poetry study [1] and the Iranian Poetry analysis [2]. These studies similarly extracted several features like in this study. Their large list of features gave inspiration for some of the features in this study.

3 Methodology

3.1 Dataset

The datasets used for training were a collection of poems known to be written by Montale and a different collection of poems written by Cima. More specifically, 115 poems from Montale and 146 poems from Cima were used. These poems would then be used to train our model to get the highest accuracy, so that when the model is used on the Posthumous Diary it would have the highest possible accuracy in attributing the poem to the correct author.

Preprocessing For preprocessing, the PDF file was converted into a text format to allow the model to manipulate the data. Before each poem's title, there was a symbol that was added manually to act as a separator. The model then split the document every time it encounters that symbol, so that every poem is on a separate line. All unnecessary symbols, such as numbers and parentheses, were also removed.

3.2 Features

After researching on other similar papers, the most prevalent features were vocabulary richness, punctuation marks, common themes, similar ending syllables, and sentence length. Since this project is in Italian and none of the researchers have any experience in Italian, some of these features had to be discarded and replaced with more generic ones. The final features used in the model were:

- Lexical richness (The complexity of the poem)
- Lexical diversity (Ratio of unique words to all words present)
- Punctuation similarity
- Polarity (How negative or positive the poem is)
- Subjectivity (How subjective the poem is)

3.3 Model

The model used for this project is an unsupervised machine learning method called K-Means. This model the points closest to a centroid and groups them into one cluster. Figure 1 is an example of how K-Means works.

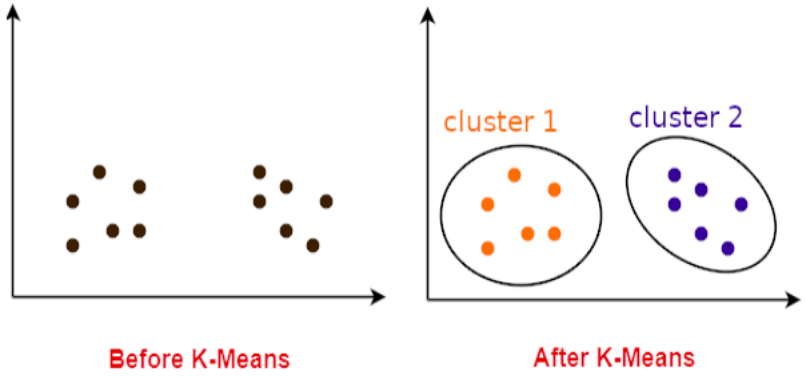

Fig. 1. How K-Means model works

In this project, we used a K-Means model with two clusters, representing the two possible cases: the poem was either written by Montale or written by Cima.

4 Results

The model was tested on the training dataset to get the most accurate version of the model. After the poems into two separate clusters, the results that came out were compared with the actual values i.e the poems that were known to be Montale's were outputted to be Montale's in most cases. This approach yielded an accuracy of 85%. This model was then used on the Posthumous Diary. The results that were outputted were 65 poems written by Montale and 28 written by Cima.

However, when comparing results to previous studies, the results varied significantly. While their results showed Montale writing 41 poems in the Posthumous Diary, this study shows that Montale wrote a larger number of poems. This difference could originate from the method the previous studies used; handwriting analysis was not used in this research. Another factor could be that the previous studies gave more weight to poems written at a date closer to Montale's death, while in this paper the poems were treated equally.

As per Table 1, the results of this study and previous studies are compared.

Table 1: Testing on Posthumous Diary

Title	Poem No.	Year Published	Old Predicted Author	New Predicted Author
Se la mosca ti avesse vista...(1971)	01	1986	AC	AC
20 gennaio o 30 anni(1971)	02	1986	EM	EM
Quando sarai imperatrice (1973)	03	1986	EM	EM
Agile messaggero eccoti(1973)	04	1986	EM	EM
È solo un vizio (1973)	05	1986	EM	AC
Ex abrupto (1975)	06	1986	EM	AC
Mattinata (1969)	07	1987	AC	EM
La foce (1969)	08	1987	EM	AC
Ma c'è chi (1973)	09	1987	EM	AC
Il clou (1977)	10	1987	AC	EM
Die Fiedermaus (1977)	11	1987	EM	EM
L'inafferrabile tua amica scrive... (1977)	12	1987	EM	EM
Mortali (1970)	13	1988	AC	AC
La congettura che il mondo... (1970)	14	1988	AC	AC
Ed ecco, nel tentativo maldestro...(1970)	15	1988	AC	AC
Al Forte (1972)	16	1988	AC	AC
Nel giardino (1976)	17	1988	AC	EM
Ricordo (1976)	18	1988	EM	AC
Nel duemila (1972)	19	1989	EM	AC
Oggi è di moda (1972)	20	1989	AC	AC
Incontro (1976)	21	1989	EM	AC
Secondo testamento (1976)	22	1989	EM	EM
L'amico ligure (1976)	23	1989	EM	EM
Come madre (1976)	24	1989	AC	EM
La felicità (1970)	25	1990	EM	EM
L'insonnia (1970)	26	1990	EM	AC
L'estate è scossa da forti temporali... (1970)	27	1990	EM	EM
Honoris causa (1970)	28	1990	AC	AC
Incertezze (1970)	29	1990	EM	EM
L'investitura (1974)	30	1990	AC	EM
King of the bay (1970)	31	1991	AC	EM
L'esegeta (1972)	32	1991	EM	EM
A sufficienza ne abbiamo di un mondo (1972)	33	1991	EM	EM
Ramo che i fortunali hanno sfrondato (1974)	34	1991	EM	AC
Settembre (1974)	35	1991	EM	EM
Tempo di distruzione (1976)	36	1991	AC	EM
Oggi un cinereo cielo grava... (1979)	37	1992	EM	EM
E' difficile vivere... (1970)	38	1992	EM	EM
Qual è la differenza... (1970)	39	1992	AC	AC
Pioggia a Venezia, neve sopra i mille... (1970)	40	1992	EM	EM
Il pesce pilota (1973)	41	1992	AC	EM
All'Onorevole-direttore (1976)	42	1992	EM	AC
Siamo burattini mossi da mani ostili... (1970)	43	1993	EM	EM
Il saggista prediletto (1975)	44	1993	AC	EM
La solitudine di gruppo... (1975)	45	1993	EM	EM
Ilritratto (1975)	46	1993	EM	EM
Telefoni per ricordarmi... (1975)	47	1993	AC	EM
Venne da me tutt'altro che sereno... (1975)	48	1993	AC	EM
Non lo sapremo mai, se furono... (1975)	49	1994	AC	EM
Il filologo (1975)	50	1994	EM	EM
All'amico editor (1975)	51	1994	AC	EM
Il ginevrino (1976)	52	1994	EM	AC
La notte che s'insinua tra le pieghe...(1976)	53	1994	EM	AC
Siccome ammiri la mia tendenza... (1976)	54	1994	AC	EM
Resta lontano dalle secche... (1974)	55	1995	AC	EM
In giorni come questi, spesso... (1974)	56	1995	AC	AC
L'amica napoletana (1974)	57	1995	AC	EM
Il criterium era scontato... (1974)	58	1995	EM	EM
Tornerà la musica che assicura...(1976)	59	1995	EM	EM
Al giovane critico genovese (1976)	60	1995	AC	AC
Sorta dall'isola che generò colombe... (1971)	61	1996	EM	EM
Non so se preferisco...(1972)	62	1996	EM	AC

Table 1: Testing on Posthumous Diary

Title	Poem No.	Year Published	Old Predicted Author	New Predicted Author
Colazione all'Augustus (1975)	63	1996	AC	EM
Il caffetano bianco(1976)	64	1996	AC	EM
Porterai con te l'ultima ventata...(1976)	65	1996	AC	EM
Ancora si crede che scrivere... (1977)	66	1996	AC	EM
Un alone che non vedi...(1969)	67	1996	EM	AC
Nell'orizzonte incerto d'una porta... (1969 su 1970)	68	1996	AC	EM
Con gli occhi fissi...(1970)	69	1996	EM	EM
S'addensano nuvole... (1972)	70	1996	AC	AC
Il tuo pallore...(1973)	71	1996	AC	AC
Un'imbeccata e via... (1974)	72	1996	AC	AC
Nel giardino dei Giusti...(1975)	73	1996	AC	AC
Nell'algida sera d'inverno... (1975)	74	1996	AC	EM
Il profumo dell'estate...(1976)	75	1996	AC	EM
Quel giorno venne Angelica... (1976)	76	1996	AC	EM
Un giorno non lontano...(1977)	77	1996	AC	EM
Deponete la vostra invidia...(1977)	78	1996	EM	AC
Per scancellare il mio ricordo... (1978)	79	1996	AC	AC
Sentivo le ore insonni...(1979)	80	1996	EM	EM
Parlerai di me con lo stesso... (1979)	81	1996	EM	AC
Vivremo mai nella nostra...	82	1996	EM	EM
Difficile è credere...	83	1996	AC	EM
Il creatore è al corrente...	84	1996	AC	AC

5 Conclusion and Future Work

Based on the results, Montale did in fact write most of the poems, but there were a significant amount of poems written by Cima as well. These results were extracted from the poems using stylometry. This is one of the possible approaches; the other being handwriting analysis. This other approach can be the method used in future research and could possibly give new insight on the Posthumous Diary. More proficient Italian speakers can also use the features that were discarded in this project to further improve the model.

6 Acknowledgements

We would like to thank the MehtA+ tutors - Haripriya Mehta, Andrea Jaba, Bhagirath Mehta, and Mohammed Sharafat - for their support and guidance. We would also like to thank Professor Paola Italia for giving us this opportunity.

References

[1] Al-Falahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. Arabic poetry authorship attribution using machine learning techniques. *Journal of Computer Science*, 15(7):1012–1021, 2019.

[2] Sohrab Rezaei and Nasim Kashanian. A stylometric analysis of iranian poets. *Theory and Practice in Language Studies*, 7(1):55, 2017.