An Information-based Model for Writing Style Analysis of Lyrics

Melesio Crespo-Sanchez[1], Edwin Aldana-Bobadilla[2], Ivan Lopez-Arevalo[1], and Alejandro Molina-Villegas[3]

[1]{melesio.crespo,ilopez}@cinvestav.mx
[2]edwyn.aldana@cinvestav.mx
[3]amolina@centrogeo.edu.mx
[1]Cinvestav - Tamaulipas, Victoria, Mexico.
[2]Conacyt - Cinvestav, Victoria, Mexico.
[3]Conacyt - Centro de Investigación en Ciencias de Información Geoespacial, Mexico.

Abstract

One of the most important parts of the song's content is the lyrics, in which authors expose feelings or thoughts that may reflect their way of seeing the world. This is perhaps the reason why modern techniques of mining text have been applied to lyrics to find semantic aspects that allow us to recognize emotions, topics, authorship among others. In this work, we focus on the analysis of syntactic aspects assuming that they are important elements to recognize patterns related to the writing style of an individual author or a musical genre. We present a theoretical information model-based in a corpus of lyrics, which allows finding discriminating elements in a writing style that could be used to estimate, for example, the authorship or musical genre of a given lyric.

1 Introduction

Text mining has been applied to the analysis of lyric content in recent years to extract valuable hidden information. Since this content is not directly amenable to numerical computation, a *feature engineering process* is applied to extract features from the text. This process may include word embeddings (Espinosa-Anke et al., 2017) or probabilistic models (McFee and Lanckriet, 2011). From a set of numerical features, it is possible to create computational models to recognize patterns associated with the content of the lyrics. This recognition allows to carry out automate tasks such as topic modeling (Devi and Saharia, 2020), semantically similar lyrics detection (Chandra et al., 2020), sentiment analysis (Akella and Moh, 2019), text summarizing (Fell et al., 2019), automatic lyric generation (Potash et al., 2015), linguistic analysis (Petrie et al., 2008), explicit content detection (Chin et al., 2018) and music recommendation systems (Dong et al., 2020). Typically, these models highlight semantic aspects associated with the content of lyrics, leaving aside other important aspects such as those associated with the way the content is written. In this regard, we propose a method that considers syntactic aspects to recognize different writing styles in song lyrics.

This work is organized as follows: In Section 2, we present the underlying ideas that support the main line of our proposal. In Section 3, we present the assessment methodology to determine the effectiveness of our proposal. In Section 4, we show obtained results from experiments. Finally, in Section 5, we present some brief conclusions.

2 Background

One important aspect when we have a text is to quantify the information in it. Where important elements of text must be identified, we considered the following concepts.

2.1 Information modeling

Information in a text can be defined as the facts about a situation, person, idea, among others, that are part of a document that follows certain grammatical and vocabulary rules in any language. This information can be modeled in the next ways:

- *Information Content*. Given a random variable Y, the information of the event $(Y = y_i)$ is inversely proportional to its likelihood. This information is denoted by $I(y_i)$ and expressed as (Shannon, 1948):

$$I(y_i) = log\left(\frac{1}{p(y_i)}\right) = -log(p(y_i)) \quad (1)$$

  For example, is the word love very common in lyrics? the informative content (I love) is expected to be low compared to other words less likely.

- *Shannon Entropy.* The expected value of I is known as Shannon's Entropy, which is de-

fined as (Shannon, 1948):

$$H(Y) = -\sum_{i=1}^{N} p(y_i) log(p(y_i)) \quad (2)$$

When $p(y_i)$ is uniformly distributed, the entropy of Y is maximal. This means that Y has the highest level of unpredictability and, thus, the maximal information content. Regarding the lyrics, the entropy rises when songs are more heterogeneous (in terms of grammar diversity).

2.2 POS Tagging

A key task in the text analysis is the Part Of Speech Tagging (POS Tagging). The POS of a word is its grammatical category associated. From the categories, it is possible to find grammatical structures and recognize elements related to things, ideas, people, etc. relevant to the information in the text. The POS Tagging is the process in which words are marked in a document as their corresponding POS category, based on its definition and context. A word strongly depends on its context and may have different categories depending on it (Toutanova et al., 2004).

3 Proposal

Our proposal aims to model the writing style of artists from a corpus of lyrics. In a corpus of lyrics grouped by artist, we encoded for each artist, the lyrics as a set of codes that represent syntactic structures based on POS (see Section 2.2). These codes are denoted as a discrete random variable Y_i. The distribution of Y_i is approximated through the frequency of the codes in the artist's lyrics. Based on this distribution, the entropy $H(Y_i)$ can be computed. We hypothesize that these values could represent a measure of *the diversity of grammatical structures* contained in the discourse of the lyrics. Based on these ideas, we propose a method that follows the pipeline illustrated in Figure 1.

As mentioned, we have a song lyrics corpus description shown in Section 4. We analyze the way artists use syntactic structures in songs to determine how are written, we focus on the POS tags used and their combinations (syntactic structures). The description of each process is described as follows:

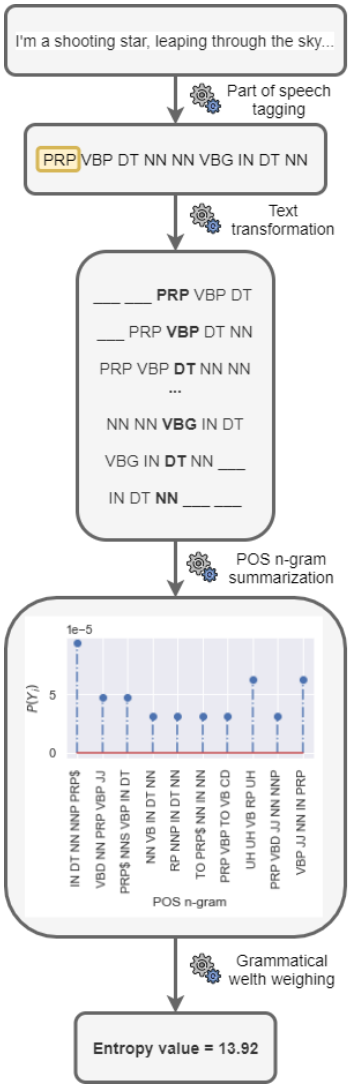

Figure 1: Proposed method to obtain an Information-based Model for Writing Style Analysis of Lyrics.

1. **Pre-processing.** Punctuation symbols, line breaks, and extra spaces removal are the pre-processing tasks in this process.

2. **Text transformation.** We extract the syntactical structures from the lyrics as follows: From the original text, a list L of POS tags is obtained. For each $l_i \in L$, we define a POS n-gram as a concatenation of l_i and its surrounding POS tags, expressed as: $l_{i-w} + ... + l_{i-1} + l_i + l_{i+1} + ... + l_{i+w}$, where w is the number of left and right POS tags, implying a length of the n-gram of $n = 2w + 1$. The above concatenation represents what we call *grammatical structures* used by an artist. When there are no POS tags to the left or right of l_i, we fill these spaces with underscores (_), this happens when l_i is either at the beginning

or the end of the lyrics.

3. **POS n-gram summarization.** POS n-grams can be seen as symbols used in the lyrics which stands for the random variable Y defined in Section 2.1. To summarize the syntactic structures used by the artists, we take all the POS n-grams previously obtained and model them as a probability distribution by the artist.

4. **Grammatical wealth weighing.** We intend to determine the writing style of the lyrics performed calculating the entropy of the POS n-grams distributions through Shannon's entropy function defined in Equation 2. We used log_2 in this paper. By performing this operation, we measure the variety of POS n-grams used in the lyrics. As a result, we could say that the grammatical wealth in lyrics is being weighed, where such entropy abstracts the artists' writing style.

4 Results

In the experiments, we used the 55000+ Song's Lyrics corpus obtained from the Kaggle repository[1]. This corpus is composed of around 55 thousand instances of English written lyrics with the next features: *Artist*, *Song*, *Link*, and *Text*.

We used a value of $w = 2$ in this experiment. Since there were artists with very few lyrics on the corpus, only instances from artists with more than 100 lyrics were taken into account. For the POS tagging task we used the Stanford POS Tagger (Toutanova et al., 2003), which is reported to have a token accuracy of 97.24%. As result, a total of 268 different artists were selected from the corpus, whose entropy was calculated via the proposed method and ranked in descending order. We obtained the top and bottom entropy values by artists such as are shown in Table 1 and Table 2 respectively. The full results can be found on a web repository that we refer to in what follows as *experimental repository* [2].

The top ten artists shown in Table 1 are filled by Hip Hop or rap artists which are well known to use complex grammar structures as well as a diversity of word combinations (a common feature in this music genre). It is worth noting that in the same order of entropy is Bob Dylan, who won the Nobel Prize in Literature 2016 awarded *"for having created new poetic expressions within the great American song tradition"*. Opposed to the artists using complex grammar structures is the bottom ten artists shown in Table 2 which are characterized by using simple grammar structures.

Rank	Artist	Entropy	Genre
1	LL Cool J	14.5948	Hip Hop
2	Insane Clown Posse	14.5056	Rap
3	Lil Wayne	14.4834	Hip Hop
4	Fabolous	14.4310	Hip Hop
5	Drake	14.3708	Hip Hop
6	R. Kelly	14.2982	Hip Hop
7	Kanye West	14.2623	Hip Hop
8	Bob Dylan	14.2475	Folk
9	Indigo Girls	14.1750	Rock
10	Joni Mitchell	14.1160	Jazz

Table 1: Top 10 artists ranked per entropy values, which denote the highest grammatical diversity.

Rank	Artist	Entropy	Genre
259	Warren Zevon	13.0663	Rock
260	Norah Jones	13.0592	Jazz
261	Wishbone Ash	13.0533	Rock
262	Whitesnake	13.0485	Rock
263	Regine Velasquez	13.0433	Pop
264	Misfits	13.0151	Punk
265	Steve Miller Band	12.9785	Rock
266	Yngwie Malmsteen	12.9450	Metal
267	Planetshakers	12.6079	Christian
268	Nirvana	12.4815	Rock

Table 2: Bottom 10 artists ranked per entropy values, which denote the lowest grammatical diversity.

The above can be summarized in Figure 3 wherein is shown the distribution of the entropy values per artist. The tails of the distribution contain the lower and higher entropy values, corresponding to the top and bottom artists previously shown. We can argue that an artist with a higher value of entropy, exhibits a greater diversity of POS n-grams in its lyrics than artists whose entropy is lower. This diversity, in terms of entropy, is an interesting finding that could reflect the grammatical wealth of the artists' lyrics.

Notice that the entropy values follow a normal distribution wherein the most likely values (around the mean) would correspond to artists

[1] https://www.kaggle.com/datasets
[2] http://bit.do/entropy_lyrics

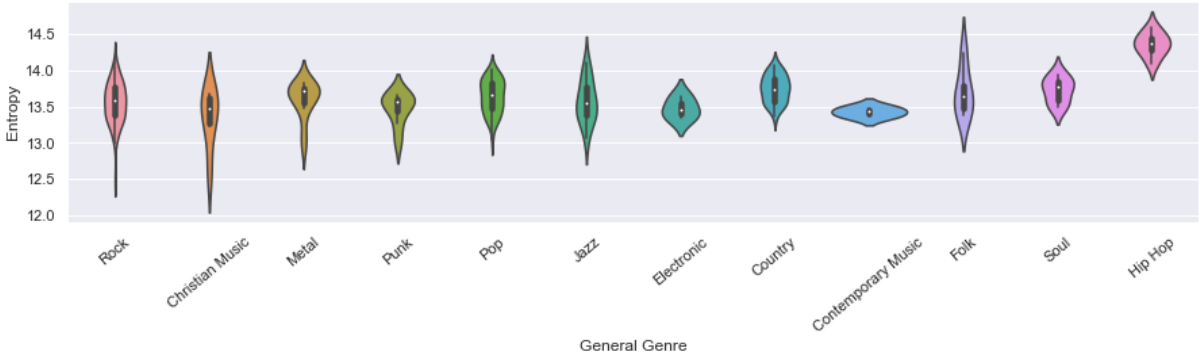

Figure 2: Artists' entropy distributions per music genre.

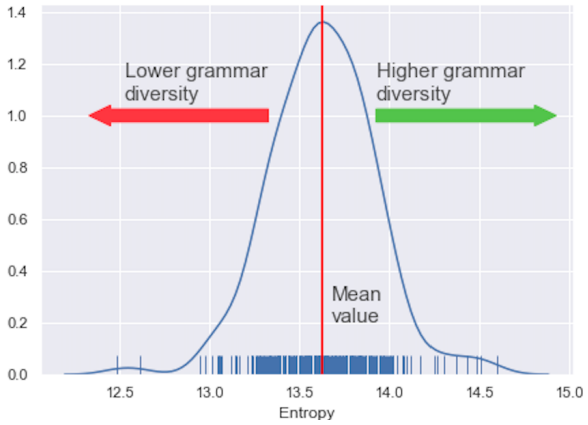

Figure 3: Artists' entropy distribution.

with an ordinary writing style (in grammar diversity diversity terms). We are interested in the tails of the distribution, where we can find the lowest and the highest diversity. In this regard, we have included in the *experimental repository*, a comparison of two lyrics corresponding to the artist with the lowest and highest entropy (*Nirvana* and *LL Cool J* respectively). We can see an important difference between these lyrics from what we have called grammar diversity.

We conducted another experiment to remark the entropy distribution per music genre. In this case, we manually labeled the 268 artists where the results of these distributions are shown in Figure 2 represents the distribution of the artists' entropy values by genre. Here the differences between genres can be appreciated in the shape of the distributions.

The variance between distributions also tells us some differences between genres. Comparing the *Rock* and *Electronic* genres. In the first one, we can find lyrics with a very low or very high diversity of POS n-grams, given the variance of the distribution. Furthermore, the language in electronic music tends to be simple because this genre tends to focus more on music than on lyrics, this effect can be resumed in the variance of its entropy distribution, which is lower than in the first case.

The entropy in music genres can be different even if they have similar variance in their distributions, such as the case of *Country* and *Hip Hop*. In this case, Hip Hop lyrics tend to have higher values of entropy than Country lyrics since they use more POS n-grams to make rhymes.

5 Conclusions

In this work, we have analyzed how artists make use of grammatical structures trying to identify the writing style in their lyrics (syntactical approach) rather than the meaning of words in them (semantic approach). By abstracting the syntactical structures that are used in the lyrics with an entropy value, we have found that the writing style in lyrics tends to approximate a normal distribution which can be the result of the common syntactic structures used in the English language. Nevertheless, remarkable observations were found in certain artists' writing style and also when analyzing entropy by music genre. We intend that the results obtained in this analysis can be used to develop a new representation that refers to lexical, semantic, and syntactic elements in the abstraction of the text.

References

Revanth Akella and Teng-Sheng Moh. 2019. Mood classification with lyrics and convnets. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 511–514. IEEE.

J. Chandra, Akshay Santhanam, and Alwin Joseph. 2020. Artificial intelligence based semantic text similarity for rap lyrics. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–5. IEEE.

Hyojin Chin, Jayong Kim, Yoonjong Kim, Jinseop Shin, and Mun Y Yi. 2018. Explicit content detection in music lyrics using machine learning. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 517–521. IEEE.

Maibam Debina Devi and Navanath Saharia. 2020. Exploiting topic modelling to classify sentiment from lyrics. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 411–423. Springer.

Yuchen Dong, Xiaotong Guo, and Yuchen Gu. 2020. Music recommendation system based on fusion deep learning models. In *Journal of Physics: Conference Series*, volume 1544, page 012029. IOP Publishing.

Luis Espinosa-Anke, Sergio Oramas, Horacio Saggion, and Xavier Serra. 2017. Elmdist: A vector space model with words and musicbrainz entities. In *European Semantic Web Conference*, pages 355–366. Springer.

Michael Fell, Elena Cabrio, Fabien Gandon, and Alain Giboin. 2019. Song lyrics summarization inspired by audio thumbnailing.

Brian McFee and Gert RG Lanckriet. 2011. The natural language of playlists. In *ISMIR*, volume 11, pages 537–541.

Keith J Petrie, James W Pennebaker, and Borge Sivertsen. 2008. Things we said today: A linguistic analysis of the beatles. *Psychology of Aesthetics, Creativity, and the Arts*, 2(4):197.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Kristina Toutanova, Dan Klein, Christoper Manning, and Yoram Singer. 2004. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—NAACL '03*, 1.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.