Generating song lyrics

Exploring Different Machine Learning Approaches To Generate Song Lyrics

Lukas Busch
iloveeatlukas@gmail.com

Sarah de Jong
sarahdejong@outlook.com

Tom Klein Tijssink
tomkatee@hotmail.com

Abstract

The goal of this project is to explore different machine learning approaches to generate song lyrics. We use a data set that contains 362.237 songs. First, we use a BERT model in order to add a positive or negative sentiment to each song. Then, we explore a simple N-gram model, a word-based LSTM, a character-based LSTM, and a GPT-2 model to generate song lyrics. After evaluating the results with a survey judged by people, it is found that the GPT-2 model performs best.

1 Introduction

Text generation is a complex task that many researchers have tried to master over the last few decades (Zock & Adorni, 1996). A topic that has not been researched much, however, is song lyrics generation. Although song lyrics are of course a form of text, there are many differences between song lyrics and for instance a book or an article. Some of these differences include the length of the sentences and the repetition of words. In this project we will, therefore, explore song lyrics generation.

2 Related work

Song lyric generation is a specific application of text generation. With works such as GPT-2 (Radford et al., 2018) and GPT-3 (Brown et al., 2020) OpenAI text generation has become increasingly effective and popular. In (2021), Duval et al. used GPT-2 to create a writing tool that can be used by both authors and artists. In (2020), Nikolov et al. created 'Rapformer', which generates rap lyrics. For their research, Niklov et al. used a BERT-based paraphrasing scheme, which yielded impressive results. One can see how state-of-the-art language models are leveraged to generate song lyrics.

3 Data set collection and description

We use a csv file data set that contains 362.237 songs from kaggle, which can be found *here*. The data set contains 362.237 songs, which includes the song name, artist, year, genre, and lyric. The songs are from 18.231 different artists, 52 different years, and 12 different genres.

4 Methods and main algorithms

In this section, we will discuss the methods and algorithms we used. Note that since we are limited in the length of this paper, we will refer to other sources for some of the specific definitions and details.

4.1 Adding Sentiments

For this project, we wanted to include another parameter besides song name, artist, year, genre, and lyric, that we could use to differentiate between the style of a song, as we hypothesized that the number of samples might become too little when separating the data based on artists and that separating data per year would not have a big enough impact. For this reason we introduced the 'sentiment' parameter, which can either be *positive* or *negative*. To generate the labels per song we use the following simple approach:

- First, we train a BERT classification model (Devlin, Chang, Lee, & Toutanova, 2019) using a data set that contains annotated sentiments for lines of poetry (Sheng & Uthus, 2020). In this data set, each sentence had a score of -1 (negative); 0 (neutral); 1 (positive), or 2 (undefined). We use the training/validation and testing data set s as provided by the article and train a BERT model that received an F1 score of roughly 85%, which was similar to the authors' results (Sheng & Uthus, 2020),

- Then, we apply the BERT model to each line of a song lyric. This way a song is now expressed as a list of sentiment scores. After filtering out the *meaningless* scores of 2, we simply sum up the list to define the sentiment score for that song. If the sum of scores is higher than 0 we label the song *positive*, if the sum is lower than 0 we label it as *negative*.

We realize that 'the sentiment of a song' is an abstract and ambiguous statement and we feel the need to clarify that our predicted 'sentiments' do not fully represent the emotional message of a song. This is due to the limitations of our simple model, and the fact that our model only focuses on the lyrics, and does not take the music into consideration.

4.2 Basic N-gram model

The simplest model we implemented was a basic N-gram model. We implemented this model using trigrams. More information on N-grams can be found in an introductory book on NLP/Machine Learning, such as the one by Eisenstein (2018). We generate lyrics from the probability distribution obtained with this model by using top-k sampling, which will be discussed in Section 4.5.

4.3 LSTM model

In order to improve on our first model, we decided to make two LSTM models. The words an LSTM model predicts can be vastly different from a basic N-gram model because an LSTM model uses its previous state (the one used to predict the previous word) to predict the next one. This is done to account for the vanishing gradient problem, which in our case effectively means that the further a word is back in a sentence, the less impact it has on the current word generation. This can mean that the perceived topic of a sentence can be vastly different at the end of said sentence than it was at the beginning. A visualization of an LSTM cell can be found in Section 4.3. More information on LSTM models can again be found in an introductory book on NLP/Machine Learning, such as the one by Eisenstein (2018).

To keep things relatively comparable, one of the LSTM models is trained on trigrams much like the basic N-gram model. Besides training an LSTM model using word-tokens as the input, we also trained a character-based model. We split up our text into chunks of 30 characters and trained our model to predict the next character. We trained both LSTM models roughly the same amount of time. We again generate lyrics from the probability distribution obtained with these model by using top-k sampling, which will be discussed in Section 4.5.

4.4 GPT-2 model

To generate song lyrics using the GPT-2 model, we made use of the *gpt-2-simple* library, which is a project by MIT and allows for easy and quick interaction with GPT-2 models (Woolf, 2021). We used the *small* 124M model and fine-tuned it once on positive songs and once on negative songs.

4.5 Text generation

After obtaining the probability distributions from either the basic N-gram model or the LSTM models, we used top-k sampling to generate text. As discussed by Holtzman et al. (2020), *top-k sampling* is a method that has recently seen an increase in popularity. It takes the top k next possible tokens / characters and samples based on their probability. The important part is choosing a proper value for k. Picking a low k might generate plain and boring text, but picking a value that is too high will most likely result in nonsensical texts. For the basic N-gram model, we generated text with a k-value of 15. Since the basic N-gram model generates a probability distribution based on words, there are often multiple *correct* next tokens. With similar reasoning we picked a k-value of **15** for the word-based LSTM model. For the character-based LSTM model, however, we picked a k of 5. This is because the number of next possible characters that result in a sensible text is much lower than the number of possible words.

5 Results and findings

A positive and negative song generated by the basic N-gram model can be found in Appendix A, a positive and negative song generated by the word-based and character-bases LSTM models can be found in Appendix B, and a positive and negative song generated by the GPT-2 model can be found in Appendix C. Of course, songs in any genre can be generated using our models, but we made the arbitrary choice to focus on pop songs. As one can see from the examples in Appendix B, the character based model is much more likely to

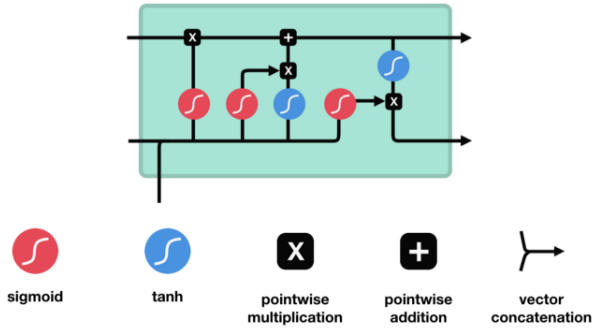

Figure 1: A visualization of an LSTM cell, Source: Towards-Data-Science

make grammatical mistakes. An option would be to lower the value of k, but this results in a boring or repetitive text.

After having generated these songs with the different models, we want to compare and evaluate them. We thought of the following ways to do this:

- Give the first sentence of a real song and see how similar the lyrics our model outputs is to the original song.
- Check the morphological correctness of the generated lyrics. We can for example check this by taking a pre-made dictionary and checking if the words in our generated lyrics are in this dictionary, in order to make sure that our model is not producing gibberish.
- Check the syntactic correctness of the generated lyrics, for example by using a parser.
- Check by humans.

We will not use the first method, since the fact that the generated lyrics is not similar to the original song, does not necessarily mean it is a bad song. Since most of our models were trained using word-tokens, the second approach also does not really make sense as a means of comparison. The third approach would be useful when for example generating the text of a book, but not for generating lyrics, since lyrics are often not syntactically correct, because they include sentences such as "yeah yeah yeah". Therefore, we have decided to use the last approach. We created a survey which can be found *here*. The survey includes 9 songs: one positive and one negative pop song generated by each of the 4 models, and one pop song written by a person. The song written by a person is "I want you" by Marvin Gaye, which we retrieved from *here*, and can also be found in Appendix D. The songs were all generated using the first words "I want", which is why they all start with these words. The survey includes three questions:

Model	Average points	P/N guessed	P/N true
N-gram (E)	1.83	10/2	P
N-gram (A)	1.83	2/10	N
LSTM words(D)	1.5	4/8	P
LSTM words (G)	1.25	7/5	N
LSTM char (F)	1	2/10	P
LSTM char (H)	1.83	1/11	N
GPT2 (I)	3.5	10/2	P
GPT2 (B)	3.08	3/9	N
REAL (C)	4	9/3	X

Table 1: Survey Results for generated songs

- To what extent does each song sound as a song written by a person? (1 being definitely not written by a person and 5 being definitely written by a person)
- All songs have been generated to be pop songs and are either 'positive' or 'negative'. How would you classify each song?
- How would you rank the songs? (write as e.g. "ABCDEFGHI")

We received twelve responses. Although this is a relatively small amount, it is enough to get an impression of how people judge our songs. A summary of the results can be found in Table 1, which includes the model, the average number of points for question 1, the amount of times a song was guessed to be positive and negative, and the 'true' positive or negative sentiment, as assigned by the model from Section 4.1. The results of the first question can be found in Figure 2.

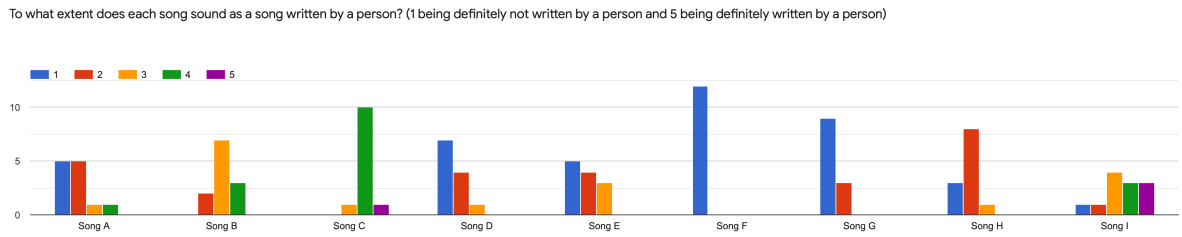

Figure 2: Ratings for "written by a person" from the survey

6 Conclusion and discussion

6.1 Analysis of the survey results

Except for the songs generated by GPT-2 and the real song, the songs are rated quite low. Out of our models, we see that the GPT-2 model receives the highest rating, then the simple N-gram model, and then the LSTM models. There were 4 people that said the GPT-2 positive song is the best song in the last question, so better than the real song. We also see that the GPT-2 positive song received more ratings of 5 than the real song. Surprisingly, the word-based LSTM positive song is rated higher than the character-based LSTM positive song, while the word-based LSTM negative song is rated lower than the character-based LSTM negative song with a significant difference. It is also interesting to see that the real Marvin Gaye song only received a score of 4 when people were asked to judge whether it was written by a person.

Something one may notice when looking at Table 1, is that the lyrics generated by the basic N-gram model and GPT-2 model were able to capture the sentiment well enough for most people to guess correctly. This surprised us as we took a simple approach in adding the sentiments and are not even sure ourselves what 'the sentiment' of a song means. However, this term is still a bit ambiguous, since songs are not always either positive or negative. We also see that there was no unanimous vote on the sentiments of the real song.

One person commented when ordering the songs: "Very difficult, as I didn't 'understand' one of them. Those that sounded/felt most like a pop song appealed most to me. Those with weird words or weird metaphors appealed the least." This brings up an interesting point. The first step of generating a good song, is to make it grammatically correct and readable (especially the LSTM-character based model had trouble generating correct text). The second step is to make it understandable and make you feel something, as many songs written by humans do. This second step is much more difficult to achieve.

6.2 Limitations of our project

There are multiple limitations to our research. First of all, due to our choice of models, not all of our models are directly comparable as some are character-based while others are word-based. Therefore, it is hard to say whether is was the model type that made the difference when comparing a character-based LSTM model and a word-based basic N-gram model. Similarly, having used character-based models also means that a model can make spelling errors, which by human standard, brings the chance of a song being written by a person drastically down.

Another limitation of our research shows itself when looking at the the survey results. In the real song (song C) the most voted and average rating for the chance of the song being written by a person was 4. This implies that the participants of the survey cannot fully discern a real song and might be more likely to score all songs lower with their preconceived notion that the songs in the survey are not written by a person (since we asked them to evaluate songs we generated using machine learning models, and did not tell them there was a real song written by a person in the survey as well).

Furthermore, our subject matter of song lyrics lends itself to limitations, as a model trained on language misses out on learning nuances such as rhythm, which negatively impact its credibility.

Lastly, it is hard to draw any solid conclusions from participant's sentiment estimates as the sentiment labels are added by our own model and any discrepancies between a label and a participant sentiment guess may also be due to the limits of our sentiment model.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. Retrieved from https://arxiv.org/abs/2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Duval, A., de Leseleuc de Kerouara, G., & Lamson, T. (2021). Controllable and contextualised writing tool for novel authors. *ArXiv*, *abs/2101.03216*.

Eisenstein, J. (2018). *Natural language processing.* MIT Press, Georgia Tech, USA.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). *The curious case of neural text degeneration.*

Nikolov, N. I., Malmi, E., Northcutt, C. G., & Parisi, L. (2020). *Rapformer: Conditional rap lyrics generation with denoising autoencoders.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Sheng, E., & Uthus, D. (2020). *Investigating societal biases in a poetry composition system.*

Towards-Data-Science. (2018). *illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation.* Retrieved from https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Woolf, M. (2021). *gpt-2-simple.* https://github.com/minimaxir/gpt-2-simple. GitHub.

Zock, M., & Adorni, G. (1996). Introduction. In G. Adorni & M. Zock (Eds.), *Trends in natural language generation an artificial intelligence perspective* (pp. 1–16). Berlin, Heidelberg: Springer Berlin Heidelberg.

Appendix A

Positive simple N-gram model:
i want to know
i know i need you to stay in shape check the vibe is right
just give me just for today
i'll be right
so i don't you give it up
and we'd be nowhere
and i'm sorry
so much more that he'll find a way of what i have a thing
that my love to you
i wanna stay with you
i was your age
in the sky
i'm gonna be a friend who's so much love when we'll see me with love
no matter what they say it's what i know that i want to go
it's what's got to be
and you can't want you to love me
and you were here
you don't get enough
i'm in the night away
when the going gets rough ;
the one girl
all that i'll be with you
i'd like to love you baby

Negative simple N-gram model:
i want to let go of time
i'm gonna get it girl
i am a poor woman living on a sunny day you'll be there
when you'd be the one i care for you
when you call me
you can be

you're going down
get it get down
when the world is full of
my first love again
so why don't you think your tears when she dance
when you see me cry
you know what you want it want to
you don't let the good times are hard to forget you !'
so if you don't you want me
i'll never see you
i'ma need a friend

he's a good girl down girl
and i am
you can't let you go hmm
you've got my hands
and you'll have to let it go
i'm a freak of the day that i'm not gonna sleep tonight

Appendix B

Positive word-based LSTM model:
i want it baby loves me
somebody i can be that thing baby is hard else i can do and cry
please in love in love of all
making your eyes
i can't get your heart was sorry
if i see and if it is i got a reason to tell go
waking how is your eyes is giving no chance so much
and when i need to let your arms baby
and it's my little shoe
but won't be proud with the love where your way for the world is
worth
oh all we were standing
waiting for
gotta day has always come out to all it can tell your life for the b-side

Negative LSTM word-based model:
i want to feel so i know you want him my fool
you don't do the things that i'm good
with on me i love a million man to a playa'
'cause man oh oh for me oh will find another questions
if you wanna take it from my own part about the hint
baby from a shelf thinkin a trois
show to what i do
give you will take a grip
baby on a heart of here
huh to i feel on the front is my blue part is fame
wonderin' what she ever them? leny thong never go me
c'mon in love when they see to be so strong
my love you

Positive LSTM character-based model:
i wants when the sun that 's going in my hands to take me
the same wish you 're all arart ,
in the mishsy
i can do it
and i 'm there , we saw the collice
a girls , the rain , tonight i 'm always stole you and it 's too the maling on a man and sate the morning
i could try to get the thought
and i will not be saying the seeoess
the way would take you in my mind
they 're gonna get over
i was standing to my mind ,
in a shade , the manic time
the sad count of the words that i 'm somewhere i 'm gone
well , you got you to think that i stay
sell you , this can 't stop and the soul
and when you are the only risenenty
and i 'm not loving me , to should ceen waitin ' me
and i will never love my song ,
it 's this time that i will break i was giving

Negative LSTM character-based model:
i want to stand up ,
and she 'll be without you
it coulders in him a close tears
we wanna stay
we said , so long
this is whah , when i was a womin
arte tonight i 'm a gan of the day
she 's gating it too
i don 't have to see that you want to late to hit
you got your love i should be me ,
tell me that if you want me
you can 't leave heard of my life
when i could break me and you
and when we 're around ,
i 'm gettin ' to me
i don 't wanna stay
all one to me ,
i 've gadt it , it 's not through the way you was money
the sign is with a much or nothing
where is a shadows of the stars .
there 's the world is gonna give you
there 's no surre more
i can 't see you ,
and we would late it taking a seal

Appendix C

Positive GPT-2 model:
I want so bad now, pack my bags, take a ride
Talking, listening, glancing, talking
Flirting with everyone
Talkin', talking, all the time
And then at some point in the morning
You hear somebody on the radio scream
They scream, scream, more
Hey, hey, hey, hey, hey
They scream, scream, louder
And then, years later
Pray to the sky
To the sky
Oh, oh, oh, oh
Together, together
In time
Up and down, together
Love, loveEnough got that all wrong spun the wheel, put people on the track
Like, babe, you, babe, you,
cause you,
pounce in the room, the time goes fast
You hit the road, the end, the final tab
They swear you're gonna change, oh, oh, oh
When you get your place, go, go, go
Come on, go, work 'til the wee hours, lalalalala
Hey 'til the wee hours, lalalalalala
They say you know what you're up to, before you go up to the bars
Know what you're up

Negative GPT-2 model:
I want to burn with you
I want to burn without any for the love that you give to me
You're the one I wish for
I want to get on the ground
I mean it
(So be it)
I'm mine and I'm yours
I can't stand to lose it
I want to burn with you
I want to burn without any for the love that you give to me
You're the one I wish forThis my day and I just wanted to let go to live another day
I'm letting go of all my worries and all of my hopes
A year left after giving up the one thing I would love
Knowing my promise I had left me all at the table
And I can hide away its not the place to be
It's just a matter of leaving me and not being around
And I can hide everything from the pain it takes
Just to look into my eyes and not to see as much
I'm living the thought that I do to let it stand
Knowing my promises I left me all at the table
And I can hide away its not the place to be
It's just a matter of leaving me and not being around

Appendix D

Marvin Gaye song: *I want you uh, uh, uh*
I want you, the right way
I want you
But I want you to want me, too (want me, too)
Want you to want me, baby (want you to want me, baby)
Just like I want you
Ooh-ooh, ooh
Oh, I'll give you all the love
I want in return, sweet darling
But half the love is all I feel
Ooh, it's too bad, it's just too sad
You don't want me now
But I'm gonna change your mind
Someway, somehow, aw baby
I want you, the right way
I want you
But I want you to want me, too
I want you to want me, baby
Just like I want you
Ooh-ooh-ooh
This one-way love is just a fantasy, oh sugar
To share is precious, pure and fair, ooh
Don't play with somethin'
You should cherish for life, oh, baby
Don't you want to care
Baby, lonely out there
I want you, I want you, baby (the right way)
I want you
But I want you to want me too
Oh-oh, want you to get down, baby
When I get down with you
Yeah, darlin'
Oh, get down, baby, ooh lord have mercy
Listen precious, I want you a lot of times