Final Project Report for CS 175, Winter 2020

Project Title: Exploration of YouTube Comment Generation with RNNs

Jimmy Bi, 42583400, jjbi@uci.edu

Flavio Quintero, 45664600, flavioq@uci.edu

Kurtis Chow, 28719675, kurtishc@uci.edu

Project Summary

The project explores language modeling applied specifically to YouTube comments (text to text). Specifically, using an existing corpus of comments as input to generate individual, unique comments as output. The algorithms used to implement language modeling consist of Markov chains, word-embedding, subword-embedding, and character-embedding RNNs. RNNs were the main focus of the project. Each model was trained on a dataset of 691 thousand comments and evaluated using user studies and natural language metrics. To better gauge the effectiveness of each model, the models were also trained and evaluated on different categorical subsets of the overall dataset, specifically Music, News, and Gaming.

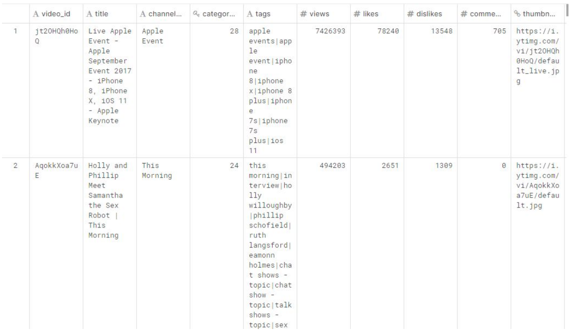
Through user studies, we compared the effectiveness of each model by evaluating how well users could determine if a comment was real or generated. It was found that among the models, the WordRNN was the best performing, able to fool participants for about 60% of the generated comments. The WordRNN was followed by the SubwordRNN, Markov Chain model, and CharRNN (hyperparameters were not controlled amongst RNN implementations). It can be concluded that for the task of generating YouTube comments, vector embedding RNNs outperform syntax based language modeling (Markov) and one-hot encoded embeddings (CharRNN); however, the task of generating casual comment/tweet-like text remains unsolved.

Data

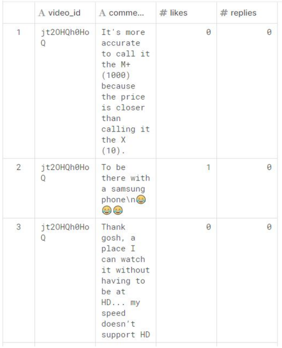
The dataset of comments along with relevant metadata was collected in 2017, containing 6 months of trending YouTube videos and their respective comments, consolidated in a 71 MB file through the use of YouTube's API. The data is presented uncleaned, in a csv file. The dataset was chosen based on the number of comments included to assist in increasing the accuracy and consistency of our language models. Collecting more recent, varied data using YouTube's API for a different corpus was explored and abandoned to save time as a result of daily API usage quotas that would reduce data collection speed.

Corpus (after preprocessing)	**Number of Characters**	**Number of Unique Characters**	**Number of Words**	**Number of Unique Words**	**Number of Comments**
All comments	56,970,623	99	12,109,323	207,568	691,375
Gaming	576,725	93	121,889	11,094	6,728
Music	7,409,607	98	1,642,250	59,035	116,295
News	5,365,285	98	1,101,777	39,329	45,866

Kaggle dataset URL: https://www.kaggle.com/datasnaek/youtube



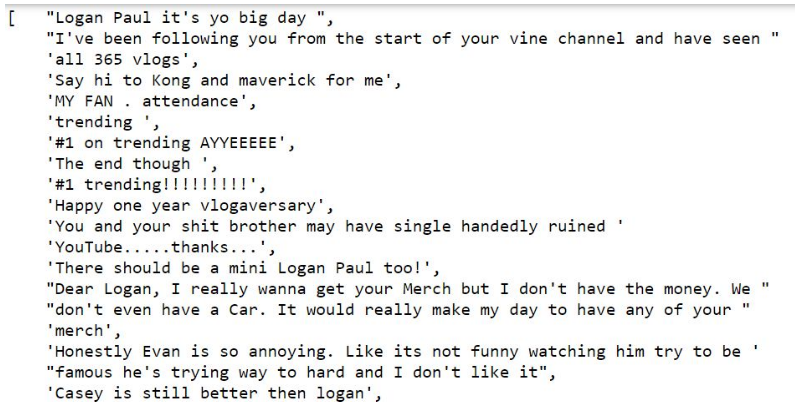
	video_id	title	channel...	categor...	tags	views	likes	dislikes	comme...	thumbn...
1	jt2OHQh0HoQ	Live Apple Event - Apple September Event 2017 - iPhone 8, iPhone X, iOS 11 - Apple Keynote	Apple Event	28	apple events|apple event|iphone 8|iphone x|iphone 8 plus|iphone 7s|iphone 7s plus|ios 11	7426393	78240	13548	705	https://i.ytimg.com/vi/jt2OHQh0HoQ/default_live.jpg
2	AqokkXoa7uE	Holly and Phillip Meet Samantha the Sex Robot | This Morning	This Morning	24	this morning|interview|holly willoughby|phillip schofield|ruth langsford|eamonn holmes|chat shows - topic|chat show - topic|talk shows - topic|sex	494203	2651	1309	0	https://i.ytimg.com/vi/AqokkXoa7uE/default.jpg



	video_id	comme...	likes	replies
1	jt2OHQh0HoQ	It's more accurate to call it the M+ (1000) because the price is closer than calling it the X (10).	0	0
2	jt2OHQh0HoQ	To be there with a samsung phone\n😂😂😂	1	0
3	jt2OHQh0HoQ	Thank gosh, a place I can watch it without having to be at HD... my speed doesn't support HD	0	0

'GBvideos.csv'

'GBcomments.csv'



```
[   "Logan Paul it's yo big day ",
    "I've been following you from the start of your vine channel and have seen "
    'all 365 vlogs',
    'Say hi to Kong and maverick for me',
    'MY FAN . attendance',
    'trending ',
    '#1 on trending AYYEEEEE',
    'The end though ',
    '#1 trending!!!!!!!!!',
    'Happy one year vlogaversary',
    'You and your shit brother may have single handedly ruined '
    'YouTube.....thanks...',
    'There should be a mini Logan Paul too!',
    "Dear Logan, I really wanna get your Merch but I don't have the money. We "
    "don't even have a Car. It would really make my day to have any of your "
    'merch',
    'Honestly Evan is so annoying. Like its not funny watching him try to be '
    "famous he's trying way to hard and I don't like it",
    'Casey is still better then logan',
```

’comments.pkl’

Figure 1: Sample Data

Technical Approach

Character, word, and subword based RNNs through PyTorch are the primary algorithms with LSTM for the hidden layer. FastText was utilized as the word embeddings for the word and subword based RNN. Input, hidden, and output layer sizes of each respective RNN reflect the size of its embeddings. For character level, input and output layer sizes consist of the number of characters (one hot encoding). For word and subword level, they depend on vocabulary size and word vector sizes.

Subword and word level RNNs differ in the tokenization methods. In the word level RNN, nltk’s TweetTokenizer for casual online text analysis was used while in the subword RNN, the sentencepiece library was used to split each of the words into separate subwords. Information like spacing was instead preserved through leading delimiter characters attached to the subwords that preceded the words. For example, “hello” to “_he” and “llo”. This in theory decreases the overall vocabulary size which in turn makes it necessary to train the model more vigorously.

Each of the RNNs were built off of a standard pytorch implementation of a bidirectional LSTM. A bidirectional model was chosen in order to preserve sentence context in both directions when generating text and LSTM was chosen over alternatives like GRU due to higher level of performance when it comes to language modeling.

Markov chains were selected as the base non deep learning model. In particular, Markovify’s implementation was selected. In this model, Markov chains are created on the given corpus and used for generating sentences by selecting words following the chains with some slight variation due to methods to randomize the selection process slightly. The base implementation along with the data was modified to fit a NewLine scheme rather than dividing the comment individually by punctuation to better fit grammar mistakes in YouTube comments.

As a method to combat overfitting on rare words and generic text, nucleus sampling, a method of thresholding the highest probable outcomes, was explored and implemented in sampling/prediction for text generation.

NumPy, Pandas, and Pickle were used to conduct preprocessing. The steps in preprocessing consisted of loading the dataset from .csv form into a Pandas frame, reading it into a NumPy array chunk by chunk, and removing any miscategorized videos and comments. Simultaneously, each comment was encoded and decoded to remove all non-ASCII characters using base Python libraries. The comments were dumped using Pickle for ease of loading into each model separately, and comments were categorized into separate .pkl files by metadata.

Additionally, a random training/validation split in the data was explored to compare fully trained models, as a method of evaluation. Perplexity was also researched as a possible method given that it fit our generation methods better than more corpus-based evaluation methods. However, due to time constraints and a focus on user evaluation, these methods were not fully implemented.

Generalized Pipeline:

1. Retrieve dataset from Kaggle's YouTube comments from trending videos for training (as csv).
2. Preprocess data by categorizing different training sets based on metadata. Consolidate each training dataset of comments into separate text documents converted from Unicode to ASCII as best as possible. Add in unused byte sequences to delineate comment endings.
3. Markov Chains: No more preprocessing necessary.
   Character embedding: No more preprocessing necessary.
   Word embedding: Tokenize each text document into words using NLTK and create word vectorizer representations using Gensim's FastText.
   Subword embedding: Tokenize each text document into subwords using SentencePiece and create word vectorizer representations using FastText
4. Feed training data through customized bidirectional LSTM RNN with varying layers in batches.
5. Adjust hidden and output layers based on loss using Adam and CrossEntropyLoss. Continue training until n_epochs is reached.
6. Generate evaluation text using trained RNN (or Markov Chain) using priming words/tokens/characters selected from a probability distribution of leading words/tokens/characters taken from the training data. The leading unit selected for priming depends on the RNN type (Either a character or word).
7. Use Nucleus Sampling to remove most probable results over a certain threshold.
8. Regenerate comment if found to match a comment in the training set.
9. Repeat training and text generation on partitions of the original dataset based on category (e.g. music, gaming, news).

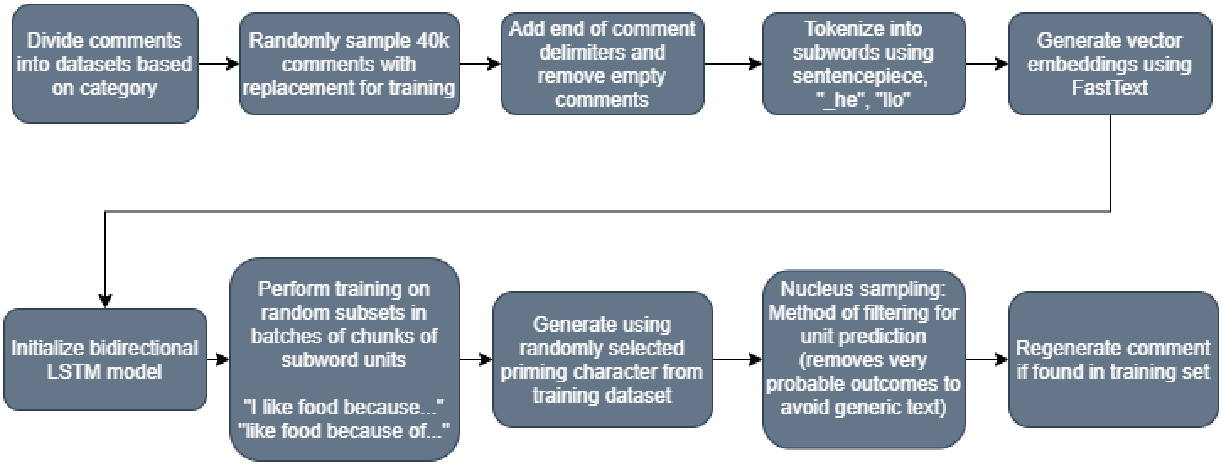
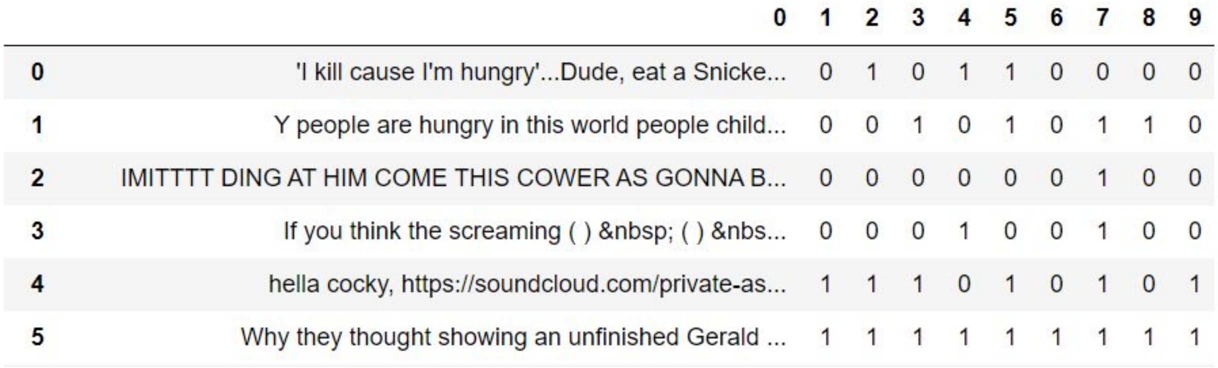

Divide comments into datasets based on category
Randomly sample 40k comments with replacement for training
Add end of comment delimiters and remove empty comments
Tokenize into subwords using sentencepiece, "_he", "llo"
Generate vector embeddings using FastText
Initialize bidirectional LSTM model
Perform training on random subsets in batches of chunks of subword units
"I like food because..." "like food because of..."
Generate using randomly selected priming character from training dataset
Nucleus sampling: Method of filtering for unit prediction (removes very probable outcomes to avoid generic text)
Regenerate comment if found in training set

Figure 2: Sample Pipeline (Subword model)

Experiments and Evaluation

Two methods: user surveys and perplexity were employed to measure model performance. The user surveys were chosen as the primary evaluation method due to difficulties evaluating short, high variance text. It is difficult to quantify performance through other metrics. One exception was made for perplexity, a measure of how well the models predict their next embedding. Data from 9 participants were gathered in total.

Surveys were set up on an application called REDCap. It was broken up into four pages based on the training data set(all, gaming, music, news) used. Each page had 20 questions: of these 20, 4 per model and 4 real comments pulled from the training set (all randomly generated or selected). Each question asked the user to read a comment and determine if they believed it to be real or generated. No feedback/results were displayed to the user.



	0	1	2	3	4	5	6	7	8	9
0	'I kill cause I'm hungry'...Dude, eat a Snicke...	0	1	0	1	1	0	0	0	0
1	Y people are hungry in this world people child...	0	0	1	0	1	0	1	1	0
2	IMITTTT DING AT HIM COME THIS COWER AS GONNA B...	0	0	0	0	0	0	1	0	0
3	If you think the screaming ( ) ( ) &nbs...	0	0	0	1	0	0	1	0	0
4	hella cocky, https://soundcloud.com/private-as...	1	1	1	0	1	0	1	0	1
5	Why they thought showing an unfinished Gerald ...	1	1	1	1	1	1	1	1	1

Figure 3: Sample survey results (1 indicates a prediction of real)

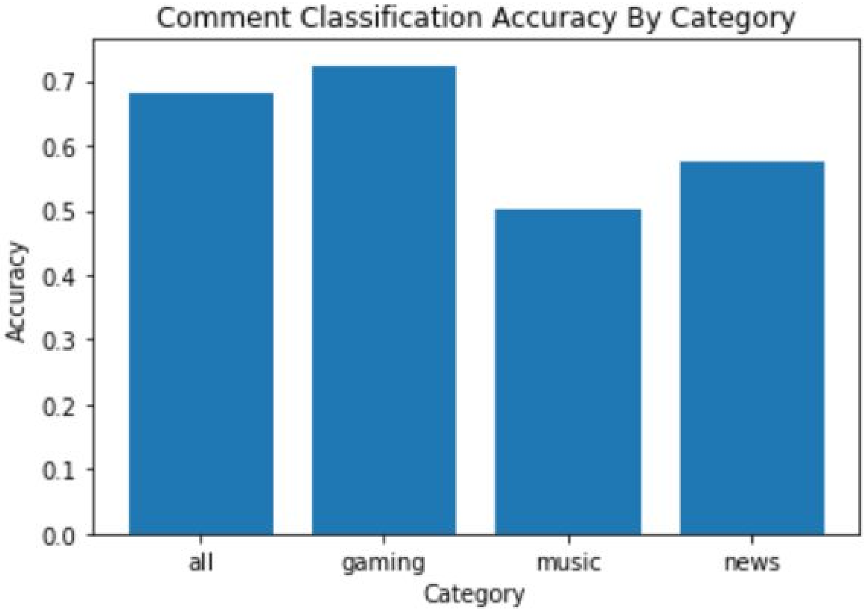


Figure 4: Comment classification accuracy by category-highest for gaming because smallest training set led to worst performing models, could be explained further by other factors

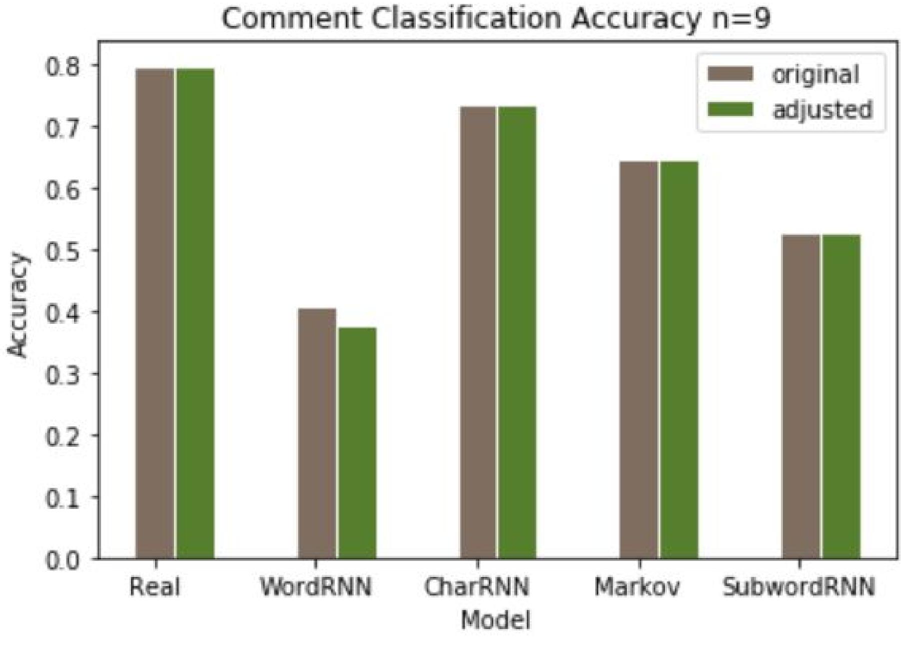


Figure 5: Comment classification accuracy by model divided by original and "adjusted" results; lower accuracy indicates participants made more incorrect guesses as to the source of the test comment. Adjusted results removes 3 WordRNN generated comments that were not postprocessed properly.

The WordRNN was the best performing model, as lower accuracy indicates that it was able to fool the most participants. However, there remains a large gap in how realistic the comments appear in comparison to actual comments taken from the training set. 1-0.799=0.201 for real compared to 0.366 for adjusted WordRNN.

On the other end, the Markov Chain and CharRNN models did not perform strongly. Comments from both models were relatively easy to recognize (0.646 and 0.736 accuracy respectively). Using the results of our models and the comments generated, the most probable explanation behind their poor performance is how their comments tended to be incoherent. For instance, in Figure 6, both comments are difficult to comprehend while also suddenly changing topic/tone in the middle.

“IMITTTT DING AT HIM COME THIS COWER AS GONNA BEASTAIDS EATEST AND AMEREDID YOU'RE THAT IT SOMETHING! This is the trailer to story thing I'm gonna come a funny, and he did a lot and laugh and we are go”
“I would tell me share it with all in hollywood that rt now its trying to replacing that thing pretty homo skulking over each computer program Joe Rogan.”

Figure 6: Top generated by CharRNN, bottom by Markov

The SubWordRNN achieved decent performance (0.528 accuracy), lower than the WordRNN despite having more training epochs. This could be attributed to the FastText generated vector embeddings containing less information (by subword instead of word). It is important to note that FastText embeddings vectorize based on subword components and not just individual words, minimizing the differences between RNN implementations to mainly vocabulary size. A larger hidden layer (vector embedding size) may have compensated for this difference and closed the performance gap.

Amongst the models, in particular the RNN’s, the common hyperparameters differed between implementations. With further optimization, performance of the models may have been improved, in particular the CharRNN. Furthermore, priming strings for the WordRNN were words, but were characters for CharRNN and SubwordRNN, which may have influenced results as leading words generated by the WordRNN were fully formed and could give more context to the comment.

Lessons Learned and Insights

We vigorously worked with RNNs to better understand the types of tasks they can perform and how to implement them in code alongside consistent modifications and additions. While they are very powerful, they are not just a black box. Be it preprocessing, tokenization, or even modifications to the RNN layers and training algorithms themselves, there are many methods of tuning the RNN to what we specifically needed to generate relevant text.

One of the surprising takeaways was how the WordRNN and SubwordRNN outperformed the CharRNN. We speculate that with the CharRNN, context and meaning are not retained as well. However, the model examines the corpus differently and rarely replicates training data.

This highlights the general strengths of the different models that we found through our experiments: context and meaning is strongly retained in WordRNN's and SubwordRNN's than CharRNN's at the trade-off of less dynamic text generation.

One of the limitations we faced during the project was our dataset. The dataset itself presented challenges inherent to YouTube comments. We could not evaluate comments on the basis of proper grammar, punctuation, and spelling, nor could we evaluate using phrases directly from the corpus. We needed a method that would assess meaning given a comment or a probability model itself. As such, we decided to focus on user testing as our primary evaluation method.

Another limitation we faced was how difficult it was to emulate the comments because of their short length and how unique they are in comparison to other comments. Although we implemented nucleus sampling as a method to address these issues, an alternative method that may have worked would be training on multiple comments at a time and generating comments in bulk (ordered chronologically or by number of likes). A grouping of comments would theoretically provide more context and display how well the model has been 'learning'.

Additionally, to be able to generate strings outside the training vocabulary, it's necessary to use pre trained embeddings in addition to the ones trained on our dataset. Selecting longer priming strings from a different set of comments not in training would help with dynamic generation.

Other limitations consisted of memory constraints (bottleneck for training vectorized word models and how long it took to train the RNN limited hyperparameter optimization) and API quotas (prevented us from doing our own data collection).

If in charge of a research lab, in order to tackle these issues, we may invest more heavily in model evaluation to compare differently tuned models, acquiring better data (YouTube comments in bulk with some ordering). It would be important to have better equipment (Google Cloud VM setups) and scheduled time along with standardized testing procedures to train/test the models. We would also try different automatic evaluation methods and compare results to user studies to find a technique that would fit our use case and measure the conservation of meaning in text generation. This would aid in tuning our models further.

APPENDICES

Appendix A: Software [at least ½ a page]

Provide a list of the major pieces of project software and their functionality (general input/output characteristics), both for (a) code you wrote, and (b) code from other people that you used. Feel free to put this information in a table if it helps to organize the information this way.

Software	**Functionality**
Markovify	Pre Existing library: Implementation of Markov Chain as baseline model
PyTorch	Library used for implementation of RNN models
SentencePiece	Library used for subword tokenization
Gensim	Library containing FastText vector model
https://github.com/spro/char-rnn.pytorch	Base implementation for the CharRNN
https://gist.github.com/thomwolf/1a5a29f6962089e871b94cbd09daf317	Nucleus sampling filtering function for text generation
Preprocessing Script	Generate .pkl of pure ASCII comments to be fed into models
Markov Chain Model	Generating text specific to markov, training and text generation
CharRNN	Training and generate scripts that take preprocessing script output as input. Training outputs a model file while generate outputs text. Uses one hot encoded characters for hidden layer.
WordRNN/SubwordRNN	Same as above, except both RNNs consolidated in a single script. Uses word vectors from FastText output as hidden layer.

Appendix B: Raw Comments (Tables Listed by Model)

Real

Category	Comment 1	Comment 2	Comment 3	Comment 4
All	'I kill cause I'm hungry'...Dude, eat a Snickers :D	Why they thought showing an unfinished Gerald in a screen test would excite... I will never know. But Cavill absolutely nails it in the show, there was always a smile on my face whenever he came on screen.	It's sad we live in the world where people make lots of money being losers with cameras, selling pictures with dumb stories for idiots to read.	DOOOOOOOOTTT!!!!!!!
Gaming	Man, this game actually makes me feel bad for robbing people. 😅	I got it! The father is a cross dresser. \J	The Canadian navy has 10 of these they use to guard the Arctic	lmao THERES the hook..
Music	This was my childhood. I watched this every time I went to my grandmother's house SD card slut..	This was so funny! Not very scary but a good idea	Sound is roadkilled	Whoa, #16 on trending
News	I really liked this video. I really disliked this video.	this is a close second. The words have such meaning and are so relateable. Cant wait to see you at the Hammersmith Apollo in March&nbsp;	this nigga highkey high af	If I was a girl, I would flash him at his stand up shows.

WordRNN (epochs=4000, layers=2, chunk_len=30, batch_size=80, hidden_nodes=300, comments=40000, iter=100, window=5)

Category	Comment 1	Comment 2	Comment 3	Comment 4
All	Y people are hungry in this world people children are homeless needing food .	I love the image of Cavill as Superman - it looks terrible hahahahaen.	T H KA HAVE .. \ nNobody died , stop with your Soros funded Vigil's and paid crisis actors !	F GET THE SECRET TO LIFE
Gaming	Spoiler for Red Dead Redemption 1* &nbsp; This boy don't make me feel nostalgic. We had one until 2007.	A Creepypasta Rituals	A black hole with the energy of a company.	B??
Music	y tal. Espero no haber ofendido a nadie, solo queria compartir lo que siento acerca de este album. Igual, lo sigo escuchando desde que salio. True Feeling, Boys on Girls y Call Me Home son algunas de mis favoritas. Que tengan un super dia.	F*K IS THIS VIDEO NOT AVAILABLE IN MY COUNTRY???*	W* >14k views &nbsp; Why is this trending?	O K you're so pretty dodie < 33
News	H.R.3999. YouTube has some informative videos on that as well. &nbsp; God bless and keep America strong.	H.R.I.P Vegas	I heard that this has something to do with your supposed profession.	s like his high school friend who died during the war, my sympathies with a nice own and bankrupted themselves. I love to slice him up slowly. Max pain. Mother fucking pedophile.

CharRNN (epochs=4000, layers=2, chunk_len=200, batch_size=100, hidden_size=300, learning_rate=0.01, temperature=0.8, predict_len=300)

Category	Comment 1	Comment 2	Comment 3	Comment 4
All	IMITTTT DING AT HIM COME THIS COWER AS GONNA BEASTAIDS EATEST AND AMEREDID YOU'RE THAT IT SOMETHING! This is the trailer to story thing I'm gonna come a funny, and he did a lot and laugh and we are go	It suggesting crop to a closet hahahaha	I'm really watching this song in this guy	I love this video for the time the drugss us up you call on youtubers easy scene right in the individams generation so this is an under my channel they losing their sisters for the worst like they're a
Gaming	I would love these thing about continuing. So make you content and also c64 and Morty's spelling most	I'm look at the exposes a much more gaming headset on it, and your videos are friendle in an example.	In the episode was manion to say this is not a simple of the same would happen but not on the best ad	I.. I just don't make a quicked it in a game? What amazing amazing turn in the title back in this vid
Music	I love her music video	I Dont, I almost like for your heart his a close to your song I listen to it, you're dead will be har	I really love the feel ourselve.	I missed us.....
News	Is rid of The Decides. I'm already to get this liability and they will be a cause the media some peop	It's better about it doesn't know what they give sales so she still all in trumps again. That haters&nbsp;	I feel a new problem is some of weeks.	It could have a fuck about like you talking about how this brace speech in anything that other weaker

Markov

Category	Comment 1	Comment 2	Comment 3	Comment 4
All	If you think the screaming ( ) &nbsp; ( ) &nbsp; ( )( ) &nbsp; ((thats fucking love TS!!!	I would tell me share it with all in hollywood that rt now its trying to replacing that thing pretty homo skulking over each computer program Joe Rogan.	I'll just like a.good rock and my life. Science supports systemic racism &nbsp; Anyone has the field,tell the board.. lol	A DOTHRAKIII HORDE NED!! ON 8MILL
Gaming	me off. You are the iron curtain was still have fun of this game ha clash of course. What safe websites can play the usefulness of rubbish	Game could either way too easily. Especially if this trailer thanks	I'm barely literate when adding micro transactions to do a fucking job oculus!	So basically generating massive open world lol. 14 + TOUCH HEY WHAT THE SWITCH! Give your talking in exchange for adding more pokemon, the Camera you're still able to do a human's impression when you try this.
Music	i love for a very boring drivel from indonesia	Yelle, you did we come together	Literally liked that i like The opening and composing her few hours. 16/10	I dont need more You gon' need a long for the name is just tell someone compared to the choreo left me is so many songs is something he's a knife?
News	What the democratic country were in the Kingdom. You really distasteful	Yeah. We are the other heads of the veteran, our veteran IT #TRUTH	YouTube	Just wait and did anything to kill, unless you given a earthquake can tell about it. Its like white and our wall....

SubwordRNN (epochs=8000, layers=2, chunk_len=80, batch_size=80, hidden_nodes=300, comments=40000, iter=200, window=7)

Category	Comment 1	Comment 2	Comment 3	Comment 4
All	hella cocky, https://soundcloud.com/private-associate-group	voice, maybe its use for my skincare... I use the it cosmetics cc cream.... I feel the exact same way as you when it comes to makeup! Not my thing at all!	falling enough the thumb pic for this vid was taken at John Travolta's sons funeral.	You are the best. You are the champion.Miracle- I am low, and I feel bad. Trump train!and have just have from Corporate come in & bully the guy making fries & see who stands up.from Brat, yea.
Gaming	no more Evening that had their density drops	fuckingn. &nbsp; Even if I knew it was a trap I'd still put it	reallycanoes erupt.	because no means the most threatening in the game but the devs gave human mains a Lol muffins!?!
Music	THIS SONG! The rhythm is so addictive, this make my day	War, Peter Parker, with the help of his way in his mentor Tony Stark, tries to balance his life as an Side note: did anybody notice the running time of this video? 4:20! Ha!	Bad Things True Blodd's theme but go metal or hard Rock on that cover. Also I think you would do amazing with Corpse Bride-Remains Of The i'm getting high on this song	was great as eldo heard this song
News	just mace like antifa and it's not our 3:24 You cant tell me Krayzie Bone didnt sound like Takeoff and Eazy E sounded like Quavo	YAHHH YAHHH YAHHH YAHHH YAHHH YAHHH YAHHH YAHHH ALL MY FRIENDS ARE DEAD PUSH ME TO MY EDGE YAHH -LIL UZI	Dhar, I often often Honestly it'll be the first to say that climate change is a real issue. I'm from Oregon and in the past few months our state has had nearly a month's worth of 90+ temperatures, some of the most	DhS PUEDE PASAR A TODOS. BESOS Y ME UNO A SU DOLOR.... Que Jehov los acompae y les de fuerzas para soportar. Y si So you're saying that tax breaks don't also help the middle class? The group that pays the most in

			devastating and Omg that was upsetting to watch	taxes and could benefit the most from a tax-cut?

Accuracy Table (n=9)

Type	Real	WordRNN	CharRNN	Markov	SubwordRNN
Original	0.799	0.410	0.736	0.646	0.528
Adjusted	0.799	0.376	0.736	0.646	0.528