

# NLP Review 1

Team 4

Afzal Mukhtar

PES2201800675

Hritika Rahul Mehta

PES2201800024

Farheen Zehra

PES2201800651

# Problem statement

Our problem statement is to “Analyse a comedian’s transcript and produce a similar script”.

Text Generation is our main goal for this project

# Literature Survey

## Paper 1

- ▶ **Published in Elsevier 2018**
- ▶ **Title** - Customizable text generation via conditional text generative adversarial network
- ▶ **Authors** - Jinyin Chen, Yangyang Wu, Chengyu Jia, Haibin Zheng, Guohan Huang
- ▶ **Description** - In this paper they propose a conditional text generative adversarial network(CTGAN) in which emotion label is taken as an input to specify the output text using variable length text generation. The CTGAN model has two modules -  $\theta$  parameterised conditional generator and  $\phi$  parameterised conditional discriminator. The initial texts generated by the CTGAN are modified to further match the real scene by an automatic word replacement strategy that extracts keywords like nouns from the training texts. To test the effectiveness of the model several datasets like the Yelp restaurant reviews and Amazon review dataset were used.

# Literature Survey

## Paper 1

- ▶ **Advantages** - There are many text generation methods which have been proposed like Markov chains and RNN but they are plagued with challenges to output a variable length text or a text different from the input sentence. The model mentioned in this paper overcomes the above hurdles. They have used two evaluation metrics which can broadly be classified into statistic based evaluation metrics and unstatistic based evaluation metrics. Based on these metrics the performance of the proposed CTGAN model is tested against other model like Markov Chain and Seq2Seq models and the given model performs better than all the other models considered.
- ▶ **Disadvantages** - When using a mixed evaluation metric to evaluate various text generators they came to a conclusion that text generated by the CTGAN is most difficult to be identified.

# Literature Survey

## Paper 2

- ▶ **Published in** ACM 2016
- ▶ **Title** - Smart Reply: Automated Response Suggestion for Email(Reference of paper 1)
- ▶ **Authors** - Anjuli Kannan, Tobias Kaufmann, Karol Kurach, Andrew Tomkins, László Lukács, Vivek Ramavajjala, Sujith Ravi, Balint Miklos, Marina Ganea, Greg Corrado, Peter Young
- ▶ **Description** - In this paper a new end-to-end approach for automatically replying to emails is described. The system proposed generates semantically diverse replies that are accessible by a single tap. This method is being used in Gmail as of now. A large scale deep learning network was used which contained LSTMs. The triggering module is the entry point into the Smart reply framework .It helps identify if an email can have short replies, a few exceptions include open ended emails and promotional emails. Data preprocessing methods include language detection(non english emails are discarded),tokenization,sentence segmentation,normalization and quotation removal.

# Literature Survey

## Paper 2

- ▶ **Advantages** - This paper introduced a new method for semantic clustering of user-generated content which requires a less amount of labelled data. Email, being one of the most popular modes of communication has led to an overload in the inbox of every individual. The given system helps suggest automated, short replies to users. Replies suggested by this model have an inherent diversity which leads to increased usage.
- ▶ **Disadvantages** - Every user sees only one response of an intent. An intent is a cluster of responses that convey a particular emotion. LSTMs which are an important component of this model have a strong tendency towards positive responses. In order to overcome this sometimes two passes of the LSTM must be done. Given that 10% of mobile responses utilise this system we can conclude that it is not widely used.

# Literature Survey

## Paper 3

- ▶ **Published in** ACM 2020
- ▶ **Title** - Diversity regularised auto encoders
- ▶ **Authors** - Hyeseon Ko, Junhyuk Lee, Jinhong Kim, Jongwuk Lee, Hyunjung Shim
- ▶ **Description** - The authors propose a new powerful text generation model in their paper, called Diversity Regularised Autoencoders. They talk about the other autoencoders that exist already and point out the issues they face. In Variation-Autoencoders has a challenge of collapsing posterior loss, which is also known as the Kullback-Leibler vanishing problem. The training is aimed at minimizing reconstruction loss. The authors have used the WGAN-GP approach to improve the training stability of the GAN and thereby improving the quality of the text generated. The noise injection strategy is applied to the encoder. The authors take four cases of noise injection - insertion, deletion, substitution, and masking.

# Literature Survey

## Paper 3

- ▶ **Advantages** - This paper modifies the existing model, which enhances the diverse nature of texts that are generated. The model maintains the readability and the grammar on noise injection, thereby it learns to correct grammar along with the training procedure too.
- ▶ **Disadvantages** - The model doesn't have any drawbacks per se, but in real life, the text is usually repeated a few times. The diversity in the model is great for learning the grammar and producing natural-looking texts, but too much diversity is not common in a normal text.



# Literature Survey

## Paper 4

- ▶ **Published in** Association for the Advancement of Artificial Intelligence (AAAI)
- ▶ **Title** - Improving Variational Encoder-Decoders in Dialogue Generation (Reference Paper)
- ▶ **Authors** - Xiaoyu Shen, Hui Su, Shuzi Niu, Vera Demberg
- ▶ **Description** - The authors have talked about improving an already present model for text generation, Variational Encoders-Decoders . They have written about the different types of models that have been used previously, and the issues they've faced. They also mentioned the KL-vanishing problem that the previous VAE models faced. In their paper, they have divided the training into two parts, the first part is to learn to encode discrete texts into continuous word embeddings, and the second part utilizes these word embeddings to learn the generalization of the latent representations. They split the model into a CVAE module and an AE module. The CVAE learns to generate the latent variables whereas the AE module builds the connection between them and the dialogue. Most of the models have a KL-vanishing problem because the RNN decoder is a universal function approximator and it tends to represent the distribution without the latent variables. The CVAE is less accurate than a GAN, theoretically as it needs to approximate the real posterior. The authors leverage a more powerful RNN encoder-decoder with this. The AE phase autoencoder the utterances to make the real posterior representable easily by the CVAE part. The experiment was performed in two datasets - the Daily dialogue and Switchboard.

# Literature Survey

## Paper 4

- ▶ **Advantages** - This model is good in maintaining the context and doesn't face issues with long sentence generation.
- ▶ **Disadvantages** - This model drops its performance if the BOW size increases. Even though it reduces the KL-vanishing problem, the performance to be in context reduces.

# Literature Survey

## Paper 5

- ▶ **Published in-** 2016 IEEE/WIC/ACM International Conference on Web Intelligence
- ▶ **Title -** Automatic Generation of Restaurant Reviews with LSTM-RNN
- ▶ **Authors -** Alberto Bartoli, Andrea De Lorenzo, Eric Medvet, Dennis Morello, Fabiano Tarlao
- ▶ **Description -** The paper demonstrates the generation of machine generated reviews which are different from genuine reviews. The paper describes related work to generate text using ANN and RNN and their disadvantages like higher computational cost. The dataset collected is composed of 2,169,264 reviews distributed over 66,700 restaurants.

They proposed a method to generate reviews given a restaurant category and rating. Their method involves 3 steps: (i) a generative phase based on a LSTM character-level recurrent neural network [They first train the network to predict the probability of the next token for a fixed-length sequence of tokens given as input—a token being a single character] (ii) a category classification phase [using Naives Bayes classifier] and (iii) a rating classification phase [pick similar sentences and assign a random R]. Two methods of evaluation used—extrinsic evaluation to assess the ability of fake reviews generated to influence the decision of a user. [Forms were set up with name, categories, reviews (at least one fake and one genuine) of the restaurant]. Intrinsic evaluation to evaluate the ability of a human user to discriminate between genuine and generated reviews. [Forms with name of restaurant and 5 reviews were tested against 39 users]. The results show that about 30% of reviews generated using their method are considered useful by human users.

# Literature Survey

## Paper 5

### ► Advantages -

- The model used is LSTM which solves the problem of long term dependencies i.e. vanishing gradient.
- The authors claim that no other paper exists for automatic generation of product reviews.
- The authors try to involve human users for evaluation who could better assess their decision than a machine.
- A diverse dataset distributed over thousands of restaurants is used.

### ► Disadvantages -

- Naive Bayes classifier ( for classification of type) indicates strong independent assumptions only- 1-grams, 2-grams and 3-grams words are considered. Instead a generative model can be used.
- The number of users considered for evaluation is less. Decisions may vary if a larger sample of users was considered.
- The authors focus only on generation of textual content of reviews.They could have considered generating reviews using other features like user activities.
- Ratings are assigned randomly to machine generated reviews.

# Demonstration

We will demonstrate the following:

- ▶ Data Gathering
- ▶ Data Cleaning
- ▶ Exploratory Data Analysis

# Contributions

## **Cleaning - Pass 1**

Afzal Mukhtar  
Farheen Zehra

## **Cleaning - Pass 2**

Afzal Mukhtar  
Hritika Rahul Mehta  
Farheen Zehra

## **Final Dataset & Visualization**

Afzal Mukhtar  
Hritika Rahul Mehta

The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect. The shapes are layered, with some appearing more prominent than others, and they extend towards the corners of the frame.

Thank you!