Contents

1. Abstract - Aditya

Social influencing is one of the biggest industries of this century and commanding control over what people have in front of their eyes is an extremely lucrative business. With all the benefits of understanding social media, there are disadvantages such as the quick viral nature of dangerous stunts and trends. Our two-fold solution is to create a multimodal model that takes in TikTok video metadata, video caption, and video thumbnail and produces a virality prediction. We used simple logistic regression, independent convolution neural networks, and finally, a multimodal neural network that amalgamates all other information into one complex network to generate the above-mentioned virality score. We successfully created simple and more convoluted networks that achieved an accuracy of 76%. The work presented here adds tremendously to a very sparse literature base that deals with social media virality in videos. Additionally, this study provides a cross-sectional analysis of currently popular deep learning and machine learning architectures that can be used to analyze the numerous components that make up a TikTok video. With our model, we could consult TikTok creators and businesses to leverage our model to provide personalized insights on growth strategies, while also working with TikTok to help them recognize dangerous trends before they become viral. We would be able to not only help small businesses and content creators expand but also ensure social media platforms remain inclusive, safe, and welcoming for all.

2. Introduction - Aditya

TikTok has over 1.3 billion monthly users and is an incredibly lucrative opportunity for individuals trying to get into social influencing or businesses looking to advertise their products. Virality is important for users aiming to attain popularity, but it also plays an important role in social influence. Trends like the Blue Whale Challenge and the Crate Challenge are dangerous trends popularized by apps like TikTok and the ability to detect and flag potentially harmful trends can go a long way in creating more safer social platforms for all.

Quantifying virality has been attempted before, but our goal is to use multimodal ensemble models and large neural nets to generate more accurate predictions. We will detail our specific approaches in the methodology section of the paper, however we were successfully able to incorporate standard machine learning techniques like logistic regression and sentiment analysis, but also more complex deep learning topics such as fully connected neural networks and CNNs.

3. Related Work - Aditya

TikTok made nearly $5 billion in 2021, so being viral on the platform is a lucrative prospect. Hence, a lot of research has already gone into curating generative, as well as predictive models, that can replicate the randomness of viral videos. Some of the relevant work is listed below.

AI-Generated Trending Video Ideas: Similar to the generative models like VAEs and GANs this project involved using RNNs to inspire the next viral video. However, they struggled with overfitting after a certain number of epochs, and gibberish titles if the model was not trained enough.

Bitgrit's 2021 Using Data Science to Predict Viral Tweets competition: While some implementations included ensemble models, most were some iteration of LightGBM or XGBoost models. This project had no deep learning component, as well as no natural language processing, however, it did a good job of highlighting preprocessing steps required for virality and how we could potentially measure virality when working with our data set.

Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks: This project involved looking at the sequential nature of posts and how they related to their popularity. This is not an attribute we are considering. Additionally, while we have ~16 features for our text analysis as well as images for CV, this research only looked at Flickr images and the views those images had, and not other important features like account size, hashtags, etc.

Additionally, the dataset we are trying to build exists on Kaggle, however, due to the nature of our project requiring us to study virality, using data from the Kaggle data set would not make sense since it is over a year old. Furthermore, the data set does not have all the features we are going to consider, as we will be using the Python TikTok API to get a fuller picture of the project.

Our project is unique since we are creating a novel data set, on a novel topic that has not been explored in the realm of TikTok videos. As one of the most popular and influential social media platforms with nearly 1.5 billion monthly active users, we must go beyond previous data sets and research. We are innovating by creating a new data set of text and images, as well as the modeling we will be doing after. Post cleaning, we will be doing a few deep learning processes, including using logistic regression on video metadata, ResNet on video thumbnails, and sentiment analysis on video captions. All of this to answer a question that has not been answered before - what makes certain TikToks more viral than others.

4. Dataset and Features - Aditya

In this section, you should focus on the description of your dataset as well as the features you're taking into account.

Since our project was on determining what makes a particular video viral, we had to make sure we had updated data from the most recent viral videos on TikTok. This meant we could not use the original dataset we found on Kaggle and had to instead learn to scrape data using the unofficial TikTok API. Since

the API only allows us to the see the top 975 videos, our dataset was of size 975 x 44 with the following columns:

'id', 'secretID', 'text', 'createTime', 'authorMeta.id', 'authorMeta.secUid', 'authorMeta.name', 'authorMeta.nickName', 'authorMeta.verified', 'authorMeta.signature', 'authorMeta.avatar', 'authorMeta.following', 'authorMeta.fans', 'authorMeta.heart', 'authorMeta.video', 'authorMeta.digg', 'musicMeta.musicId', 'musicMeta.musicName', 'musicMeta.musicAuthor', 'musicMeta.musicOriginal', 'musicMeta.musicAlbum', 'musicMeta.playUrl', 'musicMeta.coverThumb', 'musicMeta.coverMedium', 'musicMeta.coverLarge', 'musicMeta.duration', 'covers.default', 'covers.origin', 'covers.dynamic', 'webVideoUrl', 'videoUrl', 'videoUrlNoWaterMark', 'videoApiUrlNoWaterMark', 'videoMeta.height', 'videoMeta.width', 'videoMeta.duration', 'diggCount', 'shareCount', 'playCount', 'commentCount', 'downloaded', 'mentions', 'hashtags', 'effectStickers'

The most relevant columns for our purposes were:

Column Title	Explanation
'id'	Unique video ID
'text'	Video caption
'authorMeta.id'	Unique user ID
'authorMeta.verified'	Whether this user is verified or not
'authorMeta.following'	The number of accounts this user follows
'authorMeta.fans'	The number of accounts that follow this user
'authorMeta.heart'	The number of likes this user has
'authorMeta.digg'	The number of videos this user has liked
'musicMeta.musicId'	The unique music ID
'musicMeta.musicOriginal'	Whether this music is original or not
'musicMeta.duration'	Music duration
'covers.default'	Static thumbnail link
'videoMeta.height'	Height of video
'videoMeta.width'	Width of video
'videoMeta.duration'	Duration of video

'diggCount'	Number of likes on the video
'shareCount'	Number of shares of the video
'playCount'	Number of views on the video
'commentCount'	Number of comments on the video
'mentions'	Users tagged in the video
'hashtags'	Hashtags used in the video
'effectStickers'	Effects used in the video

Table 1: Column title descriptions

After doing some one-hot encoding on ‘mentions’, ‘hashtags’, and ‘effect stickers’, we obtained our intermediate dataset. After using a ratio between followers and views to determine a virality score, this was added to our dataset, finally producing the dataset that would be used throughout the ML and DL models of this project. The dataset is of size 975 x 2115.

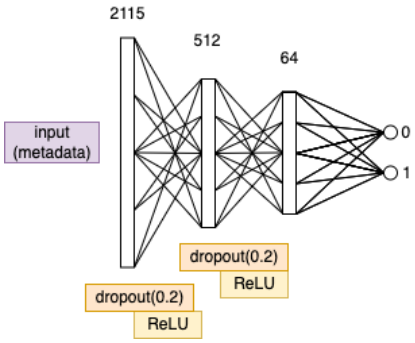
For the metadata (+NLP) analysis, we remove ‘covers. default’, as well as replace the caption with the length of the caption. Additionally, we add a new feature which is a sentiment score of the caption.

For the thumbnail analysis, we only consider ‘covers. default’, and use the static thumbnail images for the CNN and ResNet.

5. Methodology - Kelly

As mentioned in the previous section about the dataset, we have three modalities in our data: (a) metadata, (b) text, and (c) images. We first looked at creating models for each of the modalities to determine which architectures would produce a model with the best performance. By combining the best model from each modality, we constructed our multimodal model to incorporate all the modalities in our data.

5.1 Modality 1: Metadata



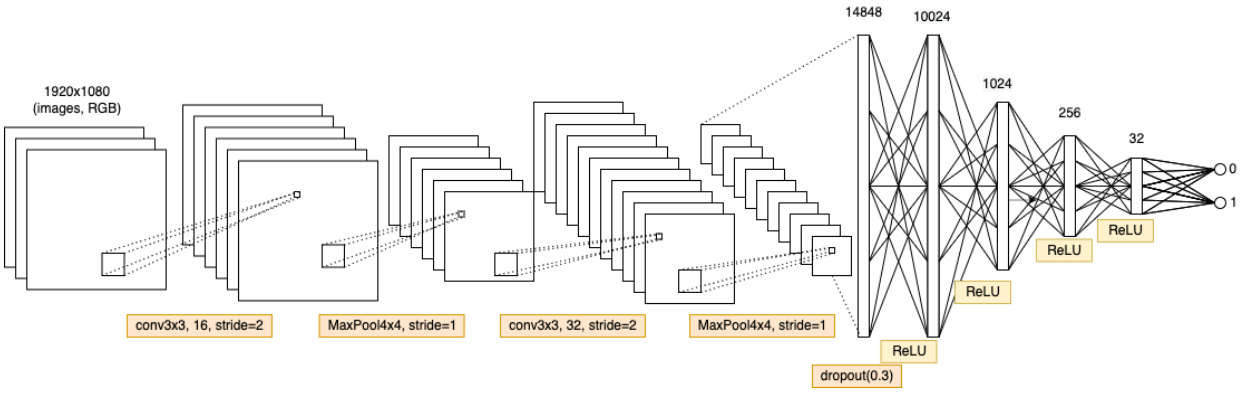
(**META_DL**)

For the metadata, we first standardized the dataset which was a crucial step in improving our performance which we had originally overlooked in our deep learning models. Our first model which we used as a benchmark was a classic LogisticRegression, created in PyTorch. To improve the performance of classification using the metadata, we created a neural network with 3 linear layers each with 512, 64, and 2 units with ReLUs as activation functions (Figure **META_DL**). Based on our training/validation loss plots, we could tell it was overfitting so we added dropouts which significantly improved our performance. Our loss function we minimized for all models was cross entropy loss and the optimizer we used was Adam.

5.2 Modality 2: Text

To extract textual information from the caption, we did sentiment analysis which produced a sentiment score for each sample. We used the function called SentimentIntensityAnalyzer from the NLTK.sentiment package to get the sentiment score. We added the sentiment score as a new feature to our metadata and used the same model architectures for the metadata to make predictions. Although there was no significant difference in performance by including the sentiment analysis, there may be relationships between the images and sentiment so we included it in our metadata for further analysis. Metadata when referenced from this point on in our methodology will include the sentiment score.

5.3 Modality 3: Images



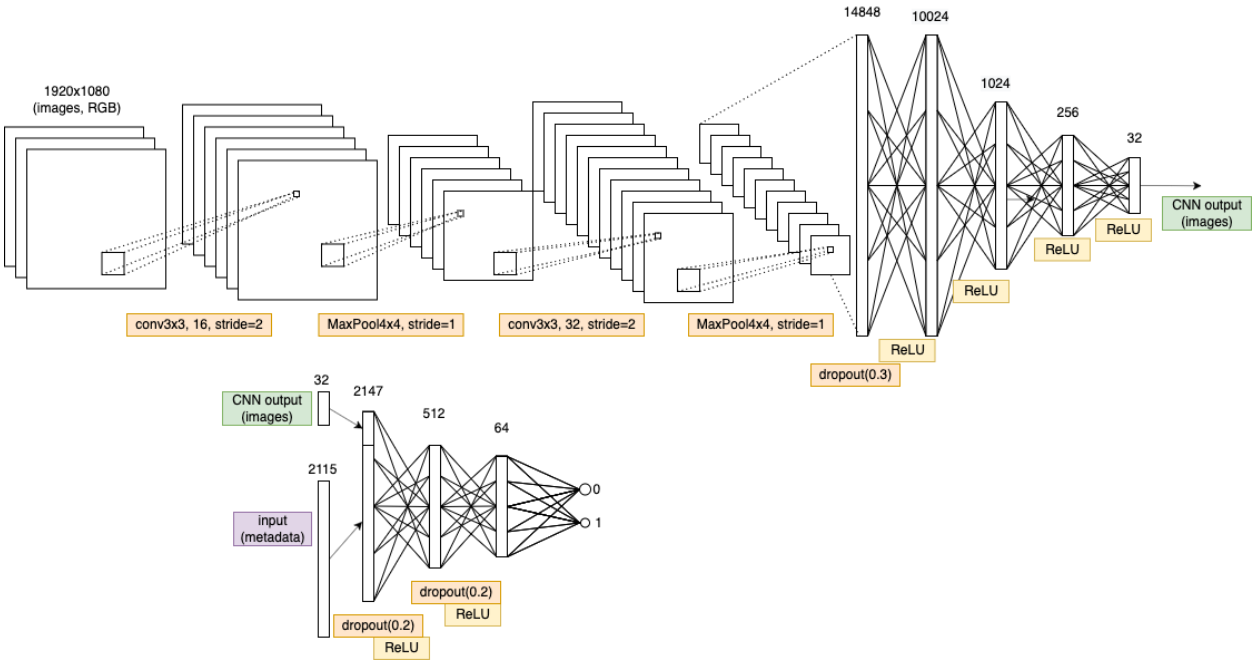
(**CNN**)

To properly extract information from images using neural networks, we needed to use convolutional layers. Our baseline deep learning model for the images was a CNN with 2 convolution layers each with max pooling followed by 5 linear layers (Figure **CNN**). Similar to the deep learning model in the metadata, we were seeing a large disparity between training and validation performance. Thus, we included a dropout layer in the first linear layer and increased variance of the images by cropping the images randomly.

Our more advanced architecture involved fine tuning the pretrained resnet50. The performance was not very different from our CNN model, but it was much more costly to train as opposed to our CNN and it

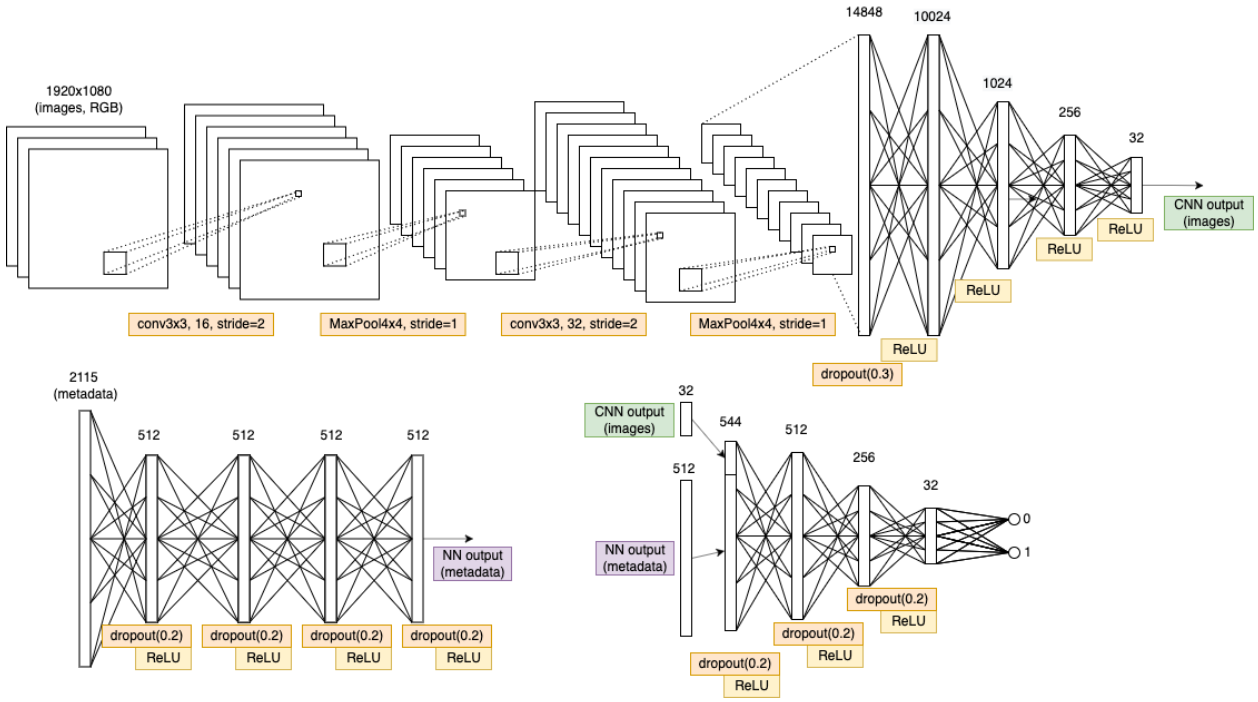
consistently crashed our GoogleColab notebook. Thus, we moved forward with our CNN for extracting important features in the images.

5.4 Multimodal: Combining the Modalities



(MM_SIMPLE) *Our multimodal architecture accounts for the different types of inputs. The CNN architecture (top) is used for the image data. The CNN output and metadata will be input into the same neural network used for the metadata from Figure* ***META_DL****.*

To bring it all together, we built an architecture that combines the CNN to the metadata to still be able to train the entire multimodal model at once (Figure ***MM_SIMPLE)***. In this overall model, we joined the features from our CNN–the last layer before the prediction output from our CNN–with the metadata as inputs into the same neural network architecture from our metadata neural network (Figure **META_DL**).



(**MM_FULL**) *Our more complex multimodal architecture differs from the original multimodal architecture in that there is a fully connected NN (bottom left) used for the metadata. Then, the model combines the two modalities to give us a final output (bottom right).*

To create a more complex deep neural network, we decided to construct a multimodal that combines our CNN from images and a new NN for metadata (Figure **MM_FULL**). The outputs from CNN and NN will be joined as inputs to a final network that includes the classification layer.

6. Results - Aditya

Start with a sentence or two on what you want to show. Describe how you showed that and what you learned. This section should include your results, i.e. the performance of your models with regards to your chosen performance metric, as well as tables/visualizations of these results, the training process, confusion matrices for classification projects… (whatever works for your project, make a sensible choice here!). Please report the loss function that you've minimized and additional measures of performance/quality you looked at, too.

The goal was to demonstrate our ability to predict viral videos. We independently tried a few models and then did a final multimodal analysis.

6.1 Modality 1: Metadata

6.1.1 Logistic Regression

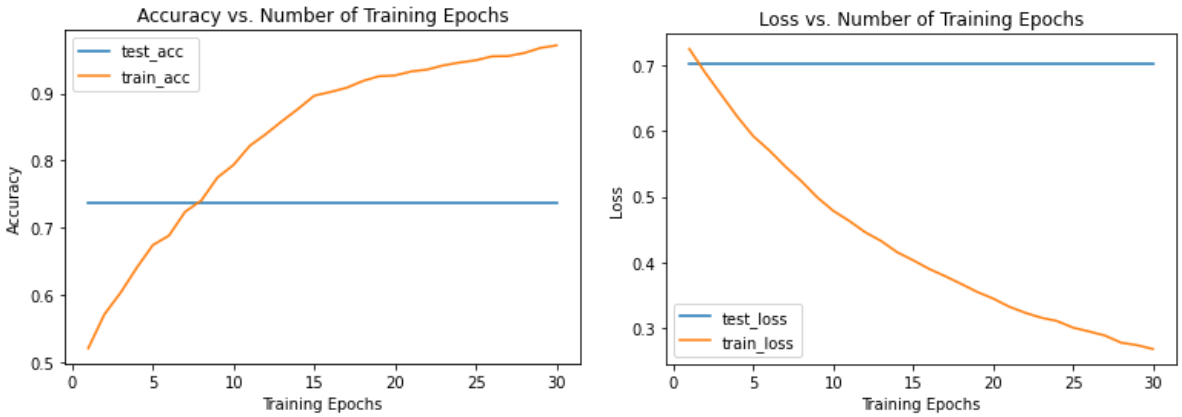


Figure 6: Accuracy and Loss vs. Number of training epochs for Logistic Regression, Metadata

From these plots, we can tell that the logistic regression as a baseline model does well in assigning virality labels based on video metadata. The test accuracy is around 75% and the loss function optimized was CrossEntropyLoss. Due to the resampling that we did prior to running these models, we do not really need to look at F1 score, and can rely on the accuracy metric alone.

	Predicted Non-Viral	**Predicted Viral**
Actual Non-Viral	87	19
Actual Viral	30	59

Figure 7: Confusion matrix for Logistic Regression, Metadata

While we don't particularly care about predicting non-viral, it is great to see that the lowest number of videos were those that were non-viral but were predicted to be viral. This shows the model is doing well at identifying viral videos and it is doing well at identifying videos that are actually not viral.

6.1.2 Deep Learning

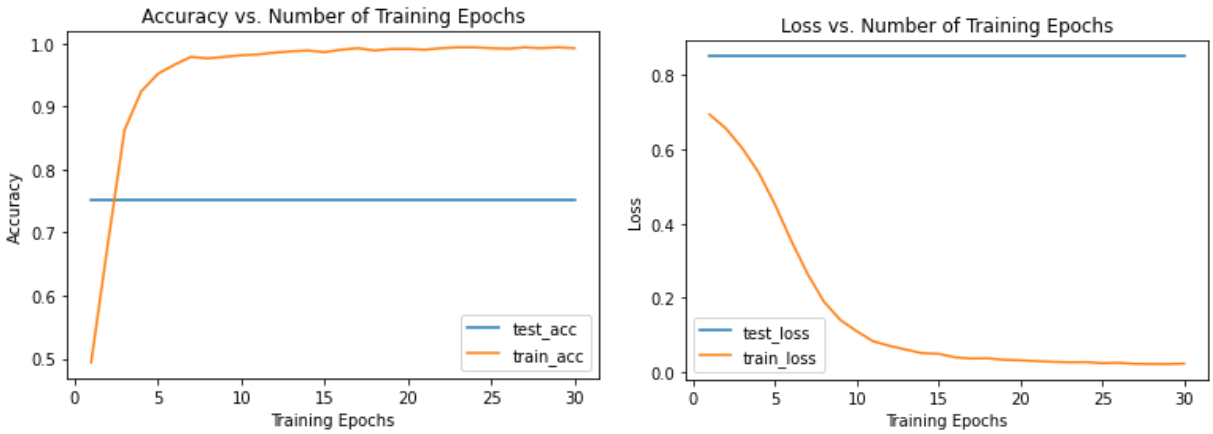


Figure 8: Accuracy and Loss vs. Number of training epochs for Deep Learning, Metadata

From these plots, we can tell that the Deep Learning model also does well in assigning virality labels based on video metadata. The test accuracy is around 74% and the loss function optimized was

CrossEntropyLoss, with the Adam optimizer. Due to the resampling that we did prior to running these models, we do not really need to look at F1 score, and can rely on the accuracy metric alone.

	Predicted Non-Viral	**Predicted Viral**
Actual Non-Viral	89	17
Actual Viral	33	56

Figure 9: Confusion matrix for Deep Learning, Metadata

The more complex model performed slightly worse, with lower accuracy for predicting viral video as viral, however it identified more non-viral video as non-viral, which is not the priority of the project.

6.2 Modality 2: Text

6.2.1 Logistic Regression

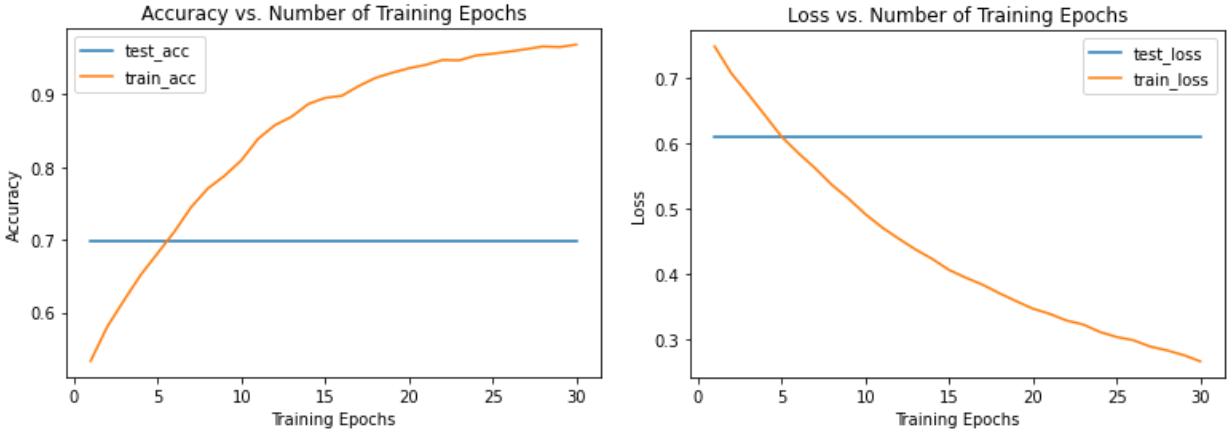


Figure 10: Accuracy and Loss vs. Number of training epochs for Logistic Regression, Metadata + NLP

From these plots, we can tell that the logistic regression as a baseline model does well in assigning virality labels based on video metadata. The test accuracy is around 70% and the loss function optimized was CrossEntropyLoss. Due to the resampling that we did prior to running these models, we do not really need to look at F1 score, and can rely on the accuracy metric alone.

	Predicted Non-Viral	**Predicted Viral**
Actual Non-Viral	79	26
Actual Viral	32	58

Figure 11: Confusion matrix for Logistic Regression, Metadata + NLP

This model performs very similarly to its non-NLP including counterpart. It does not hurt the accuracy, but it also did not completely improve the accuracy, which might suggest the tonality of the caption of a viral video might not have much to do with its virality.

6.2.2 Deep Learning

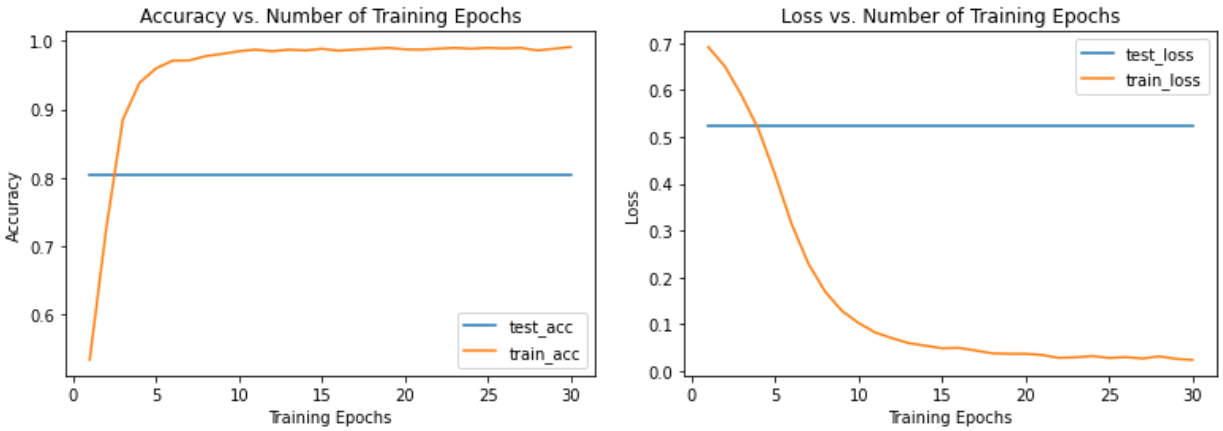


Figure 12: Accuracy and Loss vs. Number of training epochs for Deep Learning, Metadata + NLP

From these plots, we can tell that the Deep Learning model also does well in assigning virality labels based on video metadata. The test accuracy is around 77% and the loss function optimized was CrossEntropyLoss, with the Adam optimizer. Due to the resampling that we did prior to running these models, we do not really need to look at F1 score, and can rely on the accuracy metric alone.

	Predicted Non-Viral	**Predicted Viral**
Actual Non-Viral	90	15
Actual Viral	29	61

Figure 13: Confusion matrix for Deep Learning, Metadata + NLP

The more complex model performed slightly better, and performed the best from all the previous 4 models. However, there wasn't a drastic improvement, yet we decided to keep the sentiment as a feature in the metadata in case the images from the CNN are somehow correlated and a combination of the two could be a significant factor in virality.

6.3 Multimodal: Combining the Modalities

6.3.1 Deep Learning

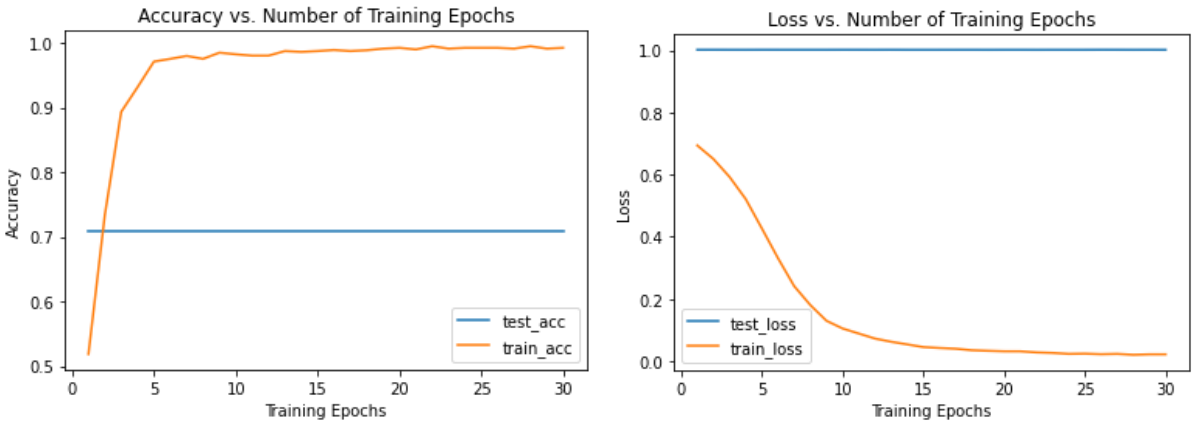


Figure 14: Accuracy and Loss vs. Number of training epochs for Deep Learning, Multimodal

From these plots, we can tell that the Deep Learning model also does well in assigning virality labels based on video metadata. The test accuracy is around 76% and the loss function optimized was CrossEntropyLoss, with the Adam optimizer. Due to the resampling that we did prior to running these models, we do not really need to look at F1 score, and can rely on the accuracy metric alone.

	Predicted Non-Viral	**Predicted Viral**
Actual Non-Viral	87	22
Actual Viral	24	62

Figure 15: Confusion matrix for Deep Learning, Multimodal

This multimodal model performed the best out of all previous models, as it is an ensemble of the deep learning models for the text, metadata and the covers. There wasn't a drastic improvement in performance, which suggests virality has little to do with the actual cover of the TikTok.

7. Discussion - Aditya

We have so far looked at three different analyses, regular meta data analysis, metadata and sentiment analysis and finally an overarching model that encapsulates text, metadata and the video thumbnails. Comparing the three models, we saw a greater accuracy in terms of the model predicting virality with the multimodal analysis.

	Predicted and Actual Viral
Metadata	59
Metadata + Sentiment Analysis	61
CNN + Metadata + Sentiment Analysis	62

Figure 16: Performance Comparison

While there wasn't a very strong indication that the multimodal was significantly better than the first two analyses (2 models each), with further training and optimization, we can potentially improve the performance

8. Findings - Raveen

Elaborate on how the performances of your non-DL, your base-DL, and your advanced models compare to each other and why one might have worked better than the others. How does your work matter? How do your results matter? What difference did you make with regards to the current research?
Include a social impact analysis here for extra credit.

MOVE SOCIAL MEDIA IMPACT TO HERE

Performance Comparison

Modality →	Metadata		Meta Data + Captions		Meta Data & Captions + Video Cover Images
Technique →	Logistic Regression	Deep Learning	Logistic Regression	Deep Learning (NLP)	Deep Learning
Accuracy →	75%	74%	70%	77%	76%

Non-DL Base Model
DL Model

Figure 17: Model Accuracies across models and modalities with test data

Comparing models

Metadata

For the metadata, our non-DL benchmark model was Logistic Regression which had an accuracy of 75%. Our DL model whose architecture is in Figure ___, had an accuracy of 74%. Originally the accuracy was much lower, but we were able to bring it up by adding dropout layers.

Captions

For the text data, we added the sentiment score produced by the text data as a feature to our metadata. We used the same models from metadata for classification which achieved an accuracy of 70% and 77% for the Logistic Regression and fully connected neural network, respectively. The reason for a higher accuracy in the DL model compared to the Logistic Regression could be due to the information learned from the interactions of the features between the captions and the metadata in the DL model. The simple Logistic Regression cannot extract information in which there are interactions between features in the dataset.

Images

Our DL benchmark model for the image data was a convolution neural network (CNN) whose architecture is in Figure _____. The accuracy was 60% which is only slightly higher than guessing randomly. However, for our advanced DL model, we fine tuned a pretrained ResNet50 and were able to achieve an accuracy of 65% of guessing virality based on just the images. Although the accuracy was slightly higher than our CNN, we used our CNN for the final multimodal model because of limited computational resources.

Multimodal

As for the final advanced DL model which encompasses all modalities of the data, we were able to only get an accuracy of 76% when predicting virality. This may be because the images do not give much information about whether a video would go viral in the first place. This hypothesis also correlates with our findings in the CNN performance for just images. In addition to this initial version of the multimodal model, we built another multimodal model which passes the metadata first through a neural network then combines with the CNN (Figure _____). The accuracy of this complex multimodal model was 46% which is worse than arbitrarily assigning labels. This may be because the complexity of the model resulted in overfitting and was not able to generalize on new unseen data.

Current Literature Overview

Most of the work done in current literature is in the realm of text-based media and its social influence. Very little research has been conducted on the impact of social media that ranges from different data

modalities such as TikTok where videos, images, and text all matter. One of the most influential and recent papers in the realm of social media influence is the work done by Leung et. al whereby the personalized DeepInf – an end-to-end framework for predicting social influence by learning a user's latent features. In this body of work, they extended the DeepInf framework by incorporating the integration of teleport probability α from the domain of page rank into the graph convolution network (GCN) model to enhance the performance of text-based data such as Twitter. Their proposed strategy has outperformed other commonly used models such as Graph Convolution Networks and Graph Attention Networks. In a similar vein, the work done by Tanvir et. al., looked into the use of machine learning and deep learning algorithms to detect fake news being circulated via Twitter. Their study did a comparison of five well-known algorithms such as Support Vector Machines (SVMs), Naïves Bayes Method, Logistic Regression, and Recurrent Neural Network models. Through their work, they were able to prove that SVMs and Naïves Bayes classifiers performed the best. This study provided insight into which features can be manipulated to ensure maximal penetration and wide acceptance of the falsely created piece of media. Another highly relevant paper is the work done by Shin et. al., which looked into the use of visual data to improve deep-learning-based social media analysis. They combine both textual components, through word embeddings, and visual components, through convoluted neural networks, to improve the performance of predicting a social media post's popularity on the Tumblr platform. However, it is important to note that the visual components used in this body of work were utilized by the neural networks to assign a complexity score to identify how well the images can hold a user's attention. The work presented in our report is a more complex implementation of the aforementioned approach with newer visual component analysis techniques on a more complicated social media platform.

Implications of the presented work

As explained in the Current Literature Overview, there is very little work done in the domain of applying deep learning to predicting social media influence and virality in social media. The majority of research done is done on single modality-based social media platforms such as text-based Twitter or image-based Tumblr. As of writing this report, there is very sparse work done on multimodal deep learning models in social media and virality prediction. Furthermore, the existing literature is focused on extracting contextual information from social media content rather than preemptively predicting the social media influence or reach a particular piece of content may have. Additionally, the work presented in this report is one of the first few large cross-sectional studies that analyze the performance of different modalities and their combinations under popular machine learning and deep learning techniques to predict social influence and virality. Consequently, the work presented here has the potential to extensively add to the current body of knowledge. Furthermore, the real-world implications of this work are detailed in the Social Implications section.

Implications of the obtained results

From our results, we can see that even in a heavily visual-based social media platform, numerical metadata and text-based caption data yield the highest accuracies when predicting the virality of a TikTok video. This provides some direction for future work to be done where more complex multimodal deep learning models can be derived with heavier emphasis based on metadata, captions, and other numerical & text-based modalities. Therefore, the results shown here provide rudimentary evidence of the immense improvement of model performance that can be obtained with further fine-tuning of the combined Meta Data & Caption deep learning model.

Social Implications

Over the years, social media has undeniably been ingrained in 21st-century society's cultural fabric. Analysis from Kepios has shown that there are "4.65 billion social media users around the world in April 2022, equating to 58.7 percent of the total global population." Within this ever-evolving space, TikTok has come to the forefront with the largest increase in planned new platform investment for brands in 2022 with an 84% increase, when compared to other platforms such as YouTube and Instagram with 66% and 64% respectively. Furthermore, social media penetration is amongst the highest in adolescents. According to Statista, 25% of all TikTok users in the US are between 10-and 19 years of age. This age group is highly susceptible to external stimuli as adolescent brain development occurs during a time where there is an imbalance between the limbic and reward systems due to the delayed maturation of the prefrontal cortex controls system – the part of the brain that is responsible for decision making and planning.

Consequently, there is a growing concern about dangerous trends propagating within the youth due to the influences of Social Media and its active user composition. Our entire team was in our late teens when the Blue Whale Challenge began and had all heard of the game. The ***Blue Whale Challenge*** involved teenagers playing a game over a 50-day period that was initially harmless, but then involved elements of self-harm and on the last day called for the participants to commit suicide. Dozens of deaths were reported from this game across the world that was popularized by social media. Alleged incidents were found in various countries including Armenia, Bangladesh, Brazil, China, Egypt, India, Iran, Italy, Israel, Russia, Saudi Arabia, and Tunisia. Similarly, the ***Milk Crate Challenge*** became popular on TikTok where participants ran over a podium-like structure being created using milk crates with both sides of the structure functioning as stairs. Due to the unstable nature of milk crates, the structure collapses during a participant's run causing numerous injuries depending on the height from which they fall. Another equally dangerous challenge was the ***Bird Box Challenge***, which also became popular on TikTok where participants would film themselves completing everyday tasks completely blindfolded, including driving. With this challenge being inspired by the Netflix thriller, "Bird Box", Netflix had to make a public statement asking fans to avoid doing the trend for their safety.

With such dangers looming on Social Media, it has become common practice within the industry to conduct content moderation to prevent the proliferation of harmful content. Given the rapid growth of social media usage across the globe, it has become harder to keep on top of all trends and viral content. Within Facebook, the company engages in a hybrid model where it utilizes algorithmic and human review processes to identify, assess and take action against harmful content. With their automated algorithmic process, they initially screen content for human moderators to make the final decision given the complexities associated with different cultural and social contexts that are ever-evolving. To further expedite the process, Facebook employs the use of digital hashes to proactively identify and block content that matches existing hash databases for child pornography and terrorism-related imagery. Given TikTok's recent surge in usage, they rely primarily on human moderators to identify harmful content as seen by their recent job postings. However, given TikTok's rapid growth, it must start employing a hybrid model like Facebook with more involvement in algorithmic models. Therefore, the work presented in this paper is a large foundational step in establishing better content moderation that can ameliorate the current laborious processes. With early warnings of which videos will become viral, human moderators at TikTok

can screen more effectively and be better equipped to keep up with the increased demand. Furthermore, Content creators can be more mindful of the content they post if they know the potential viral audience reach which can further improve the situation. Pushing these metrics to content creators is essential on TikTok more than on any other platform since TikTok's platform has the highest engagement rates in influencers with less than five thousand followers vs macro-influencers with up to one million followers. This unique characteristic incorporates an added level of complexity not seen in other platforms that make the work presented in this paper timely.

9. Limitations and Ethical Considerations - Aditya

You should describe any shortcomings of your model in this section with regards to your initial gap of filling a specific research gap. What are limitations with regards to the generalization of your findings? Furthermore, elaborate on ethical considerations, too: E.g. Could your model be misused? Are there issues with biased data that you trained/tested on?

Several limitations were imposed when conducting the aforementioned work. They are listed as follows:

1. **Data Limitations:** Due to the lack of a free TikTok scraper, we used a scraper that only allowed us to obtain 975 TikToks. This meant limited metadata, text, and more importantly, limited thumbnails. With CNNs, due to the structure of the network and the generation of different kinds of filters used, they improve drastically with a lot of images.
2. **Training Limitations:** In our project, we went with CNN instead of the ResNet due to similar accuracies and the high computational load of the ResNet. If we had more GPU power than just Google Colab, we could train both nets further and see if ResNet performed better at more epochs.
3. **Generalizations:** Due to the lack of a larger database, assumptions made in this report could be false. For example, the simpler logistic regression outperformed neural networks in some of the models described above, however, this could be a blanket generalization made because we did not have enough data or did not let the models train. This is evident from the training accuracy and loss not completely flattening out in certain above models.
4. **Biased Data:** We trained on a relatively biased data set since we obtained our data from scraping a new user's "For You" page. However, without any data about the user, this meant that a majority of the videos in our data set were viral, which isn't true for a user as they progress with using the app. They get more personalized content that can go viral, given users like the one it is shown to continue to interact with it. This was nearly impossible to model, as that would mean having to reverse engineer TikTok's recommendation engineer.
5. **Ethical Considerations:** Even if our models did not attain accuracy in the 90s, our final model was nearly 80% accurate at identifying viral videos. If our models, further improved, are used by companies to exploit our model to figure out how to make advertisements viral, it could be disadvantageous to TikTok and its users as the platform would get crowded by advertisements that were disguised as non-advertisements.

11. Future Work/Research - Raveen

Finally, think about what the next steps for your project could be if you had more time, compute power, ... Could your approach be adopted in other domains?

How could you overcome the limitations that you discussed in the previous subsection?

- Image Tagging using ImageNet
- Video Sampling
    - Context creation from video sampling
    - Generate Story associated with the clip and combine this scenario with other text based metadata that performed better than visual cues seen in the other report.
    - Find research in deep learning to caption videos

With the currently proposed system architecture, we feel that the current approach would benefit greatly with additional time to conduct more training on the data that was obtained. With our results, the team strongly believes that there is definite room to further reduce the loss across the different modalities in all the different architectural configurations. Furthermore, the team also believes that if more powerful computational hardware was available, the team could have evaluated the effectiveness of the more powerful ResNet architecture in the combined model. Lastly, with the currently proposed combined architecture, the team also believes that if more time was available, a more representative and less biased dataset could have been devised to aid with better virality prediction. Additional validation of the predictions made by the algorithm could have been done by assessing the state of the predicted videos after a certain waiting period.

In terms of architectural improvements, a viable improvement to the proposed combined algorithm is to randomly sample a few frames from each TikTok video where these images can be captioned using encoders and decoders. The captions generated can be then construed as word embeddings where the average or concatenation of these embeddings can be deemed as the “contextual” summary of the actions of the video. This data can then be passed alongside the existing metadata and caption data analysis models for a better virality prediction. Additionally, the system could also be improved by using recent improvements in video captioning deep learning models where the system can bypass the random sampling of frames done to obtain a far more accurate and comprehensive captioning of the actions present in the video that can be passed along the system architecture as presented earlier. As made abundantly clear by the work done in this report and current literature, the task of predicting virality in a highly sensory-rich medium such as video is a herculean task with many viable and varied approaches for tackling this problem.

12. Conclusions - Aditya

Our project stemmed from the importance of having a safe social media platform, but also a platform where users and content creators can understand what it means to be viral. In light of this, we conducted some research and decided to go with the most popular social media platform - TikTok. With no official scraper, we found a way to obtain a limited number of videos from TikTok and used it to do some

modeling of virality. This virality score was a ratio between a video's views and the followers that the user has. We created five models - logistic regression (LR) for metadata, deep learning (DL) model for metadata, LR for metadata with sentiment scores as features, DL for metadata with sentiment scores as features, and finally the overarching multimodal DL model. In terms of choices, we had to make some that accounted for the limitations of the project. For example, we had to choose vanilla CNN over ResNet due to ResNet being more computationally heavy. With relatively comparable performance, it was difficult to conclude one model is better than the other, but with further research, with greater computational power and more training time, a more definitive answer can be generated. One explanation is that the metadata plays the greatest role in virality, and the sentiment score, as well as the thumbnail, have little to no role in the virality of a video - there is randomness in what goes viral. The other conclusion, that we can draw from our limited data set, was that the multimodal model did perform slightly better, and this would propagate if we had a larger dataset. Hence, looking at our models and their performances, we can say while randomness might play a large factor in what goes viral, there are methodologies to understand recommendation algorithms and create models to utilize them to our advantage.