Understanding Hollywood through Dialogues

Aashna Garg, Vinaya Polamreddi

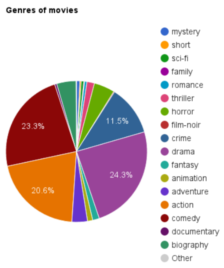
Abstract

Movies are a huge part of most of our lives. They reflect, distort and influence how our society works. The systematic bias against female and other minority characters and actors in Hollywood has been a hot topic for a while now. However, there has been very little quantitative analysis for this debate. Embodying Silicon Valley's zeal for data-driven problem solving, we explored a corpus of movie conversations and memorable quotes to computationally learn about movies and analyze the differences in movie dialogues uttered by female vs male characters using deep learning and natural language processing.

1 Introduction

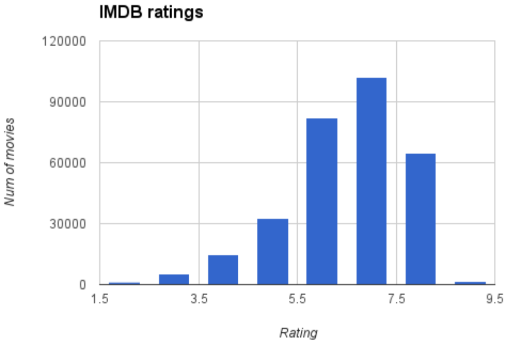
With the second year in a row of all White nominations for the Oscars with hashtags such as *#OscarsSoWhite* trending, the discrimination in Hollywood against women and minorities has been cast in the limelight. There has been a lot of media devoted to this topic, but for all the talk, there has been very little data.

The Polygraph sought out to address the lack of quantitative analysis by looking into 2000 films and attributing every spoken line to an actor in the largest analysis of scripts so far[9]. There were many useful and interesting findings that came out of the study. They found that only 22% of films had dialogue with more dialogue by female than male characters[8]. Out of the 30 modern Disney films, only 8 have gender parity of lines; even female led movies such as Mulan and Little Mermaid have less than 30% female dialogue[8].



With such frighteningly polar statistics, it is clear that there is indeed a problem in Hollywood. However, counting lines only goes so far. Do screenplay writers evening out the number of lines of female and male characters fix the problem? Alison Bechdel almost as a joke coined the Bechdel Test, a pass/fail test about gender representation[8]. If the film has two female characters talking and not about a male character, then the film passes, else it fails. While this test makes a go at a deeper understanding of film dialogue based on gender, it obviously has many limitations. Gravity, for example, is about a female character alone in space fails[8].

In our study, we want to explore the deeper differences between the dialogues written for female and male characters in film. We use datasets of movie conversations with metadata and memorable quotes to analyze how dialogue changes based on gender. We aim to see if the dialogues for female characters are less memorable as compared to their male counterparts. As a prior step to this analysis, we will also be using various classification models on the datasets to understand how well these deep learning models can understand and represent movie dialogue data. We will then use these trained classification models to help us analyze the differences in memorability of dialogues based on gender.



2 Related Work

For this project, we studied several papers in the topics of gender classification, and sentence classification, language modeling and understanding. Many papers studying gender classification of text use hand-crafted features such as n-grams and other contextual features. Studies have found that there are clear differences in writing styles based on gender in formal writing. A study done on tweets to identify gender found that using n-grams and profile information allowed them to identify gender with 77% accuracy with just text of tweets and over 90% with profile information. [16] Mukherjee and Liu achieved state of the art results of 88% accuracy in identifying gender of blog posts authors by using specific hand-crafted features such as: frequency of various parts of speech, stylistic features based on words used in their blogs, the usage of more emotionally intense words and adverbs, etc..[15]

In terms of language modeling and understanding, deep learning has in the recent years made huge advances resulting in state of the art results for classification, inference, sentiment analysis, etc.. Graves [14], Sutskever et al.[13] both made huge advances in deep learning especially using RNNs. Considering the papers together, we understand that the main takeaway is that Recurrent Neural Networks (RNNs), Long Short-Term Memory units (LSTMs) in particular, are very effective at understanding *sequences* of data. Overall, they show the effectiveness of RNNs to represent complex aspects of language and even predict them. While in principle a large enough RNN should be able to model and predict long sequences, most RNNs can't store information about past inputs for very long which causes instability when generating sequences. Long Short-term Memory (LSTM) is an RNN architecture designed to store and access information better than normal RNNs. LSTM reached state-of-the-art results in a variety of sequence processing tasks.

While RNNs and LSTMs have done very well in many language tasks, a different neural network architecture: Convolutional Neural Networks, have also proven to be effective at text classification. Simple CNNs have shown to get state of the art results on various tasks including question classification and sentiment analysis.

We plan to use these deep learning architectures on our domain specific dataset to classify movie dialogues focusing on gender classification.

3 Datasets

We use two datasets for this project.

3.1 Cornell Movie Dialogs Corpus

This dataset contains fictional conversations extracted from raw movie scripts with supporting metadata.[1] [11]

It has the following properties:

1. 220,579 conversational exchanges between 10,292 pairs of movie characters
2. 9,035 characters from 617 movies
3. 304,713 total utterances
4. Movie metadata including: genres, release year, IMDB rating, IMDB votes
5. Character metadata: gender, position on movie credits

3.2 Cornell Memorability Dataset

This dataset contains lines from roughly 1000 movies of varying genre, era, and popularity. [2] [12]

It has the following properties:

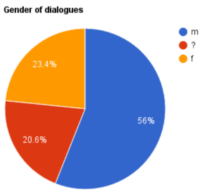
1. 894014 movie script lines from 1068 movie scripts
2. 6282 one-line memorable quotes that are automatically matched with the script line which contain them
3. 2197 one-sentence memorable quotes paired with surrounding non-memorable quotes from the same movie, spoken by the same character and containing the same number of words

4 Classification

In our first dataset, given a dialogue, we have the gender of the character, the rating and genre of the movie, etc.. We used this data to build classifiers to see if given a single dialogue in a movie, can the classifiers classify those respective characteristics. If our classifiers can learn representations of the data so that they can accurately understand the gender of a character or the rating of a movie, then it would seem that the content of these dialogues contain assertive information of other broader characteristics of the movie and character, and allows us to do further analysis.

4.1 Linear Classifier

As a baseline, we used a linear classifier, specifically an SVM. As input, we converted each dialogue into a vector by averaging a 100 dimensional pretrained GloVe vector of each word.



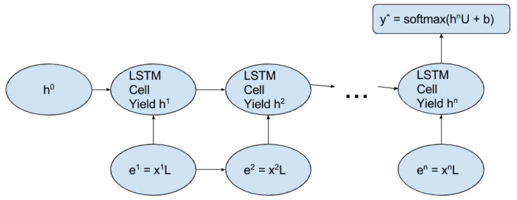
4.2 Feed Forward Neural Network

Next, we implemented a 1 hidden layer feed forward net with an additional representation layer to convert the dialogue to word embedding representation. The word embedding representation was a concatentation of the pretrained GloVe vectors for each word in the dialogue in a certain window. The size of the window was tuned as a hyperparameter.

$$\begin{aligned} x(t) &= [x_{t1}L, x_tL, x_{t+1}L] \\ h &= tanh(x(t)W + b_1) \\ y &= softmax(hU + b_2) \end{aligned} \tag{1}$$

4.3 Long Short Term Memory network

While feed forward networks have been effective in many tasks, they have many failures in understanding text. Due to the proven effectiveness of recurrent neural networks in understanding sequences including sequences of words (such as dialogues), we used an RNN as our next classifier.



We used a modification of RNN's, specifically a Long Short Term Memory network architecture. LSTM's have been proven to be better than the basic RNN architecture for most language tasks. LSTM's help capture long term dependencies in the data which is common in natural language.

The input of the LSTM was a sequence of words in each dialogue. Each word was represented as a vector using a 100 dimensional pretrained GloVe vector. The representation was then fed into an LSTM cell along with the previous timestep's hidden layer representation represented by the following equations:

$$\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{(t1)}) && \text{(Input gate)} \\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{(t1)}) && \text{(Forget gate)} \\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{(t1)}) && \text{(Output/Exposure gate)} \\
\tilde{c}_t &= \sigma(W^{(c)}x_t + U^{(c)}h_{(t1)}) && \text{(New Memory Cell)} \\
c_t &= ftc_{(t1)} + i_t\tilde{c_t} && \text{(Final memory cell)} \\
h_t &= o_t tanh(c_t)
\end{aligned} \tag{2}$$

After a certain number of time steps which was tuned as a hyperparameter, the last hidden layer representation was used as input to a softmax layer to output into the label space.

4.4 CNN

In addition to the LSTM, we also used a Convolutional Neural Network to classify our dialogues. While CNNs have generally been used in computer vision, there have been successes using CNNs to model sentences especially for sentence classification tasks. We implement a CNN model based on Kim Yoon's paper where they show a simple CNN beats many benchmarks.

In our model, we have an embedding layer where we convert our words into an embedding matrix. Then we have a convolution layer over the embedding with multiple filter sizes. After tthis, we have a max pool layer, a dropout layer and finally a softmax layer to get the output.

4.5 Discussion and Results

We ran each of the described classification models on 3 different characteristics associated with each dialogue:

1. **Memorability of a dialogue:** binary classification using data from the second dataset. The dataset consisted of 2197 memorable and nonmemorable dialogues each. The models were trained on 1500 memorable, 1500 nonmemorable quotes with a validation set of 250 each and tested on 250 of each resulting in a 3000 example train set with a 500 example data set for both validation and testing.
2. **Gender of the character speaking the dialogue:** binary classification using data from the first dataset. The dataset consisted of around 70000 female quotes and 170000 male quotes. The models were trained on 50000 male examples and 50000 female examples, 500 of each for validation set and 500 of each for test set resulting in a 100000 test set and 1000 examples each for validation and testing.
3. **IMDB rating of the movie the dialogue was present in:** 10-way classification using bucketed IMDB scores to round to closest integer. This was also performed on data using the first dataset which contains ratings. These models were trained on a random sample of 100000 examples with a validation and dev set of 1000 examples.

We will evaluate the classifications on the following metric:

1. **F1-Score** This scores measures accuracy using precision and recall. Precision is the ratio of true positives to all predicted positives. Recall is the ratio of true positives to all actual positives. Lets say the true positive is denoted by 'tp', false positive as 'fp', false negative as 'fn', precision as 'p' and recall as 'r'. The F1 score is given by:

$$F1 = \frac{2pr}{p+r} \text{ where } p = \frac{tp}{tp+fp}, r = \frac{tp}{tp+fn}$$

This metric gives equal weightage to both precision and recall and will try to maximize both precision and recall simultaneously. This would favor a moderately good performance on both over extremely good performance on one and poor performance on the other.

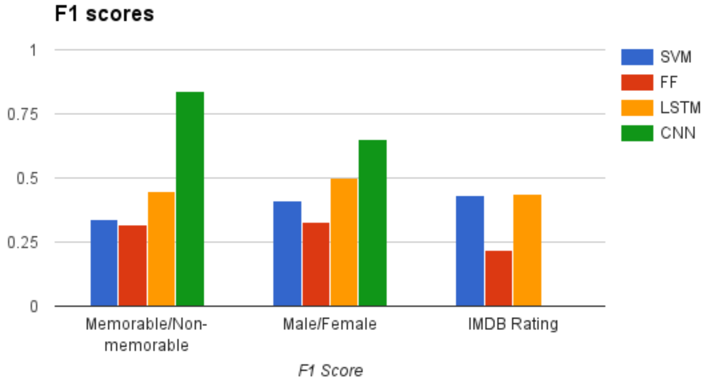


Figure 1: F1 scores on various parameters for all models

Taking a look at the F1 scores for each of the models on each classification, the Convolutional Neural Network performs better than the other three. Between the LSTM and Feed Forward networks, the LSTM consistently outperforms the Feed Forward network. SVM however does surprisingly better than the feed forward and almost as well as the LSTM in some cases. This indicates that the Convolutional Neural Network seems to have learned the best representation of the dialogues data to be able to label it with its respective characteristics.

Taking a look at the CNN loss function, we can see that even though the training loss decreases steadily, the loss for the validation stays the same and goes up near the end. This shows that the network is overfitting, and overall the labels do not seem to be classifiable by training on this data. Taking a higher look at the task, it makes sense that trying to classify the speaker's gender based on a single utterance is a really hard task especially with no longer sentences or contextual data.

5 Memorability of dialogues based on gender

As we mentioned in the introduction, we wanted to use deep learning and natural language processing for exploring the gender bias in Hollywood. There have been studies and statistics that show that women are less represented, are paid less, and given less dialogues. We wanted to know is it just quantity problem; if the script writers and casting directors even out their numbers, will the problem be solved? Or is there a qualitative difference in the types of dialogues and roles given to women and men?

The ideal way to understand if the dialogues given to men and women differ in terms of memorability is if we could analyze a dataset that has the gender and the memorability of every dialogue. However, we don't have such a dataset, but we do have a dataset of dialogues of which we know the gender and a different dataset of dialogues for which we know memorability. We propose to use the classification models previously built to classify these dialogues to get the label (either gender or memorability) that we don't have.

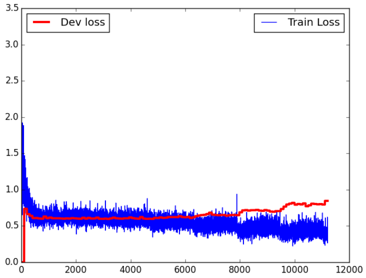


Figure 2: CNN Train and dev loss on gender classification

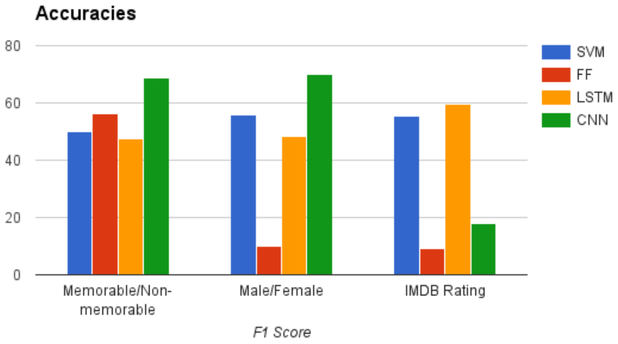


Figure 3: Classification accuracies across all models

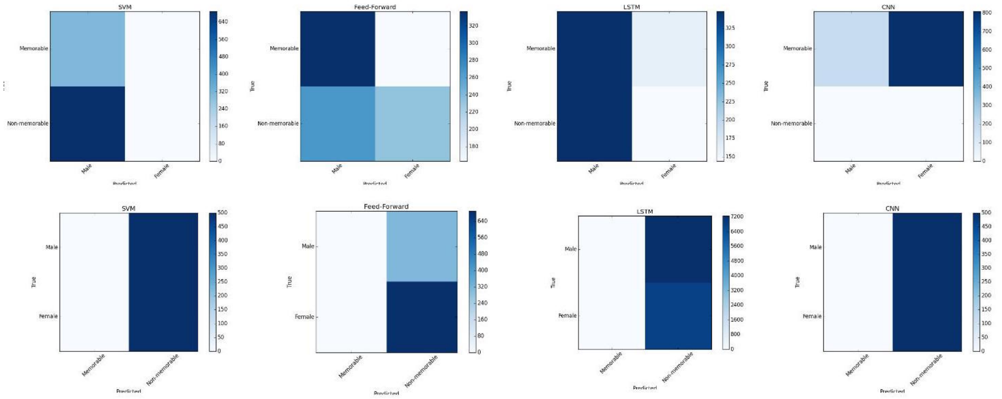


Figure 4: Gender classifer on memorability and memorability classifier on genders for different models

If we build a classifier that is good at classifying a dialogue into whether or not it looks like Hollywood's version of female spoken dialogue or male spoken dialogue, we can run a set of quotes whose memorability is known through it to get the likely gender labels and use that for analyzing the proportion of female and male dialogues that are memorable.

After we have tested various classification models described above to see if they can understand and represent data in a dialogue and if those individual dialogues actually can tell us any of the metadata associated with it, we now perform this next level of analysis.

5.1 Memorability classifiers on gender known quotes

We took the four classification models trained on memorability data and used them to classify 500 examples for which we know the gender but not the memorability.

Taking a look at the classification of gender known quotes by the CNN memorability classifier, we can see that there is no proportionality difference in how the female and male quotes were split up in terms of memorability. If we trust that the CNN classifier is an authoritative source of judging memorability, this can be used to say there is no difference in memorability of female and male quotes.

However, the classifier is not highly accurate and we cannot judge these quotes solely using any of the classifiers we have built. As a result, we propose that we analyze how the classification of the gender known quotes change as our memorability classifier become more accurate. This is based on the hypothesis that if the skew in gender/memorability skews in a definite way as the classifiers increase in accuracy, then it is highly probably that the trend actually exists. To elaborate, if the difference in female and male distribution in memorable quotes increases as the classifiers get more accurate, then it is very likely that there is a difference in the memorability of male and female quotes.

Analyzing the predictions on the gender known quotes by each of the classifiers according to accuracy, we see that the least accurate classifier according to F1 score: Feed forward, shows a larger difference in the ratio of memorable to non memorable in female and male. As the accuracy increases in the LSTM, SVM and CNN, which is the most accurate, the skew visibly decreases. This indicates that the memorability of female and male quotes is likely to be not that different.

5.2 Gender classifiers on memorability known quotes

Performing the same analysis as above but in reverse, we run our classifiers trained to predict gender of the character speaking based on the dialogue on dialogues of known memorability and unknown gender for the speaker.

It is much harder to spot any trends in the differences between male and female classified quotes as the classifiers accuracy increases. While the lowest accuracy classifier, the Feed Forward Network, shows a heavy skew, the other classifiers also show skews but in no particular trend. This is most likely due to low accuracies of all of the gender classifiers.

6 Conclusion

We used a variety of classification models to model the dialogue data to be able to classify various attributes of it. Each of the classifiers had varying degrees of success with the CNN outperforming the rest followed by the LSTM. Some labels were harder to predict than the others based on their nature. For example, predicting the IMDB rating of a movie based on a single dialogue is arguably a pretty impossible task even for humans as there are so many other factors that influence this. While gender in many other cases seems to be reasonably predictable, in this case, where we are analyzing individual utterances and long speeches or conversations and without any contextual information, we haven't been able to achieve high accuracy. We also ran multiple classifiers on the memorability dataset and while this was also a difficult characteristic to model without additional information, we achieved a reasonable F1 score with the CNN classifier.

We used these classifiers to then analyze the differences in the memorability of female and male quotes as a way to understand the qualitative gender differences in Hollywood. We performed some interesting analysis on the predictions using the classifiers and analyzed trends in predictions as the underlying classifiers increased in accuracy. We saw that when using memorability classifiers on the gender known quotes, the differences in gender became less. This seems to indicate that at least according to the memorability dataset we trained on, there do not seem to major differences in the memorability of female and male quotes. However, the reverse analysis using gender classifiers on memorability known quotes did not yield any results most likely to the overall low accuracies of the gender classifiers.

7 Future Work

There are a few directions in which this study can be furthered. Better and larger datasets of movie dialogues such as datasets with both memorability (or any other importance metric) and gender can be very helpful in training more accurate classifiers. Also being able to create better gender classifiers based on single utterance would highly increase the ability to do analyses such as described above. While many gender classifiers with high accuracies exist, they train on many hand crafted and contextual features and on longer pieces of text.

References

[1] Danescu-Niculescu-Mizil, Cristian, and Lillian Lee. ”Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.” Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, 2011.

[2] Danescu-Niculescu-Mizil, Cristian, et al. ”You had me at hello: How phrasing affects memorability.” Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012. APA

[3] Bilous, Frances R., and Robert M. Krauss. ”Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads.” Language Communication 8.3 (1988): 183-194.

[4] Bramsen, Philip, et al. ”Extracting social power relationships from natural language.” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5] Jurafsky, Dan, Rajesh Ranganath, and Dan McFarland. ”Extracting social meaning: Identifying interactional style in spoken conversation.” Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[6] Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. ”Automatically categorizing written texts by author gender.” Literary and Linguistic Computing 17.4 (2002): 401-412.

[7] Mukherjee, Arjun, and Bing Liu. ”Improving gender classification of blog authors.” Proceedings of the 2010 conference on Empirical Methods in natural Language Processing. Association for Computational Linguistics, 2010.

[8] You Can Now Search 2,000 Films for the Movies Where Women Get the Most Lines — Bitch Media.” Bitch Media. N.p., n.d. Web. 17 May 2016.

[9] ”The Largest Analysis of Film Dialogue by Gender, Ever.” Polygraph. N.p., n.d. Web. 17 May 2016.

[10] Polygraph’s Film Dialogue Dataset. Matthew Daniels, n.d. Web.¡https://github.com/matthewfdaniels/scripts/¿

[11] ”Cornell Movie-Dialogs Corpus.” Cornell Movie-Dialogs Corpus. N.p., n.d. Web. 17 May 2016.

[12] Cornell Memorability Dataset. Cristian, n.d. Web. ¡http://www.mpi-sws.org/ cristian/memorability.html¿

[13] Sutskever, I., Martens, J., and Hinton, G. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning, ICML 11, pp. 10171024, June 2011.

[14] A. Graves. Generating sequences with recurrent neural networks. In Arxiv preprint arXiv:1308.0850, 2013

[15] Arjun Mukherjee, Bing Liu. Improving Gender Classification of Blog Authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[16] John D. Burger, John Henderson, George Kim, Guido Zarrella. Discriminating Gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011

[17] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. ”Glove: Global Vectors for Word Representation.” EMNLP. Vol. 14. 2014.