Documenting Sexism in Amateur Literature: A Neural Approach

Mohammad Omar Khursheed*
Master's Student in Computer Science
University of Massachusetts, Amherst
omkhursheed@umass.edu

Tezuesh Varshney*
Senior Year Undergraduate Student, Computer Engineering
Aligarh Muslim University
tezueshvarshney@zhcet.ac.in

Ishan Singh*
Senior Year Undergraduate Student, Computer Engineering
Aligarh Muslim University
ishansingh@zhcet.ac.in

Abstract

We attempt to build a natural language processing model that detects instances of sexism in amateur writing in the 21st century, focusing on social media networking site Reddit.com's popular subreddit, r/WritingPrompts. The purpose of this study is to find trends of sexism in amateur fiction, and look to relate them to popular culture of the 21st century. The approach to documenting sexism is one that uses natural language processing techniques inspired by a class of artificial intelligence algorithms called neural networks.

The content under consideration is based upon the assumption that everyday sexism in all eras has constantly been observed to have influenced the literature of that era. Since formally published literature is edited heavily and by experts, overt displays of sexism, xenophobia, racism, etc. are often pruned, even if commonplace in real life, but in amateur fiction on the internet, such instances are left in place by authors, especially given the cloak of anonymity that the online world provides. Natural language processing in the modern technological era has taken the world by storm. With the advent of neural-network influenced natural language processing methods in areas varying from machine translation to text summarization, as well as the availability of vast amounts of data on the internet, the capabilities of artificial intelligence-based systems, such as the aforementioned neural networks, are bordering on the near-human level. We used transfer learning to derive meaning from sequences of words to determine if the text is sexist or not, which allows the model to avoid learning from scratch.

Our goal is to classify a piece of writing as either being sexist or not, and we believe that with such a system in place, it will be possible for us to look at trends in sexism over the past few years in terms of amateur literature, and perhaps correlate them with changes in political atmosphere across the world.

Keywords: natural language processing, neural networks, sexist language, classification, data science for social good

1 Introduction

In today's technology-driven world, a microcosm of society at large is obviously the internet, and more particularly, social media. Hence, analysing the actions of people on social media, where they are often veiled by anonymity unavailable to them in real life, is an interesting path that may lead to results where people's raw and unfiltered emotions are the primary source from which their commentary is derived, and hence which provides a clearer picture for us to look at in terms of societal behaviours. It is obvious, from the amount of sharply polarized opinions and debates that social media provokes, that such a study would be well proportioned to understand the trends of extreme opinions on a variety of issues that are often controversial and prone to attracting people with such opinions.

When we speak of social media, the world's traffic is concentrated on only a few sites, mostly Facebook, Instagram, etc. But when we look at topical content, Twitter is where the vast majority of immediate reactions take place, and Twitter is often seen as much more of a 'news aggregator' than the likes of Facebook are. Hence, mining text data from Twitter provides a rich and current source of information for measuring social trends in the current socio-political scenario. And in terms of Natural Language Processing, Twitter has, especially over the past few years, become extremely fertile ground for mining a lot of valuable data for a variety of purposes. A very popular area of NLP is sentiment analysis, and a Twitter data is often used for this purpose.

In our paper, though we choose to focus on amateur fiction and underlying trends of sexism in such fiction, and for this purpose, we build the main component of our system using Reddit data, specifically the data from the subreddit r/WritingPrompts[1], where people respond to prompts with short stories. We build a system that begins from using Twitter data to create a model of what sexism in the modern Internet era looks like, and then we extend that system to amateur fiction through the data we collect from Reddit.

2 Research Methodology

In our research on mapping trends in sexism in amatuer literature, we use a variety of data science and machine learning[2] technologies, and requisite background information is given in the Background section below. The following steps were taken as part of the research methodology we used to undertake this study:

2.1 Data collection

There are of course two parts to the data collection steps in this project, since we initially use Twitter data, followed by amateur fiction sourced from Reddit data. A short description of both kinds of data follows.

- **Twitter data:** Thanks to Twitter's extremely robust API[3], we were able to source Tweets using the Tweepy library[4] for the programming language Python. We sourced a labeled dataset containing tweets with labels such as `racism`, `sexism` and `none` w.r.t tweet IDs, and edited it to divide the data into two categories, sexist and non-sexist.
- **Reddit data:** Reddit also provides support to get raw data from its threads and comments section. To scrape data from subreddit r/WritingPrompts [1], we used the PRAW (Python Reddit API Wrapper) and with the Python Pushshift.io API Wrapper for comments and submission search. We filtered out the top comments from each thread, preprocessed those comments and determined the trend shift using our model.

2.2 Preprocessing

Since machine learning algorithms do not play well with raw data, we performed certain pre-processing steps. These were as follows:

- **Removal of non-textual data and links:** This is especially important in the case of Twitter data, since the characters `@` and `#` are extremely popular on the social networking site, as are links to various web pages which are not useful for the purpose of this study.

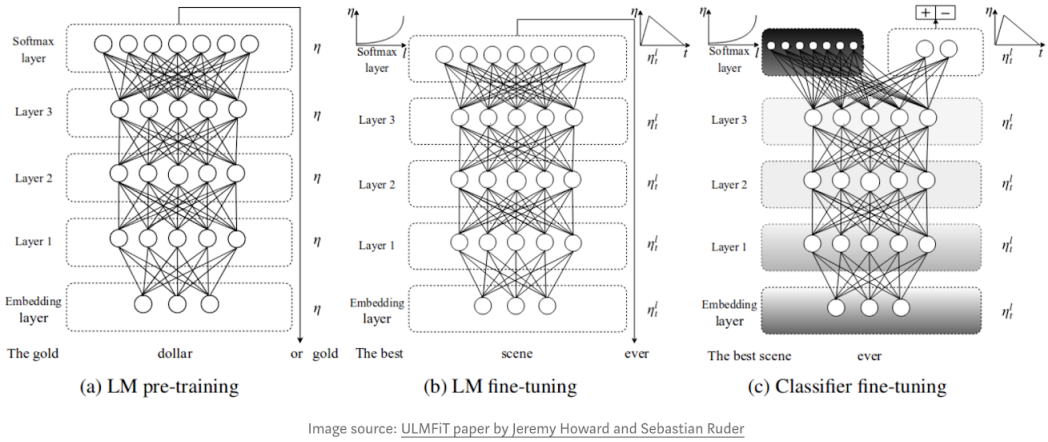

Image source: ULMFiT paper by Jeremy Howard and Sebastian Ruder

Figure 1: Three steps involved in ULMFiT

- **Tokenization and stop word removal:** Stop word removal and tokenization are standard preprocessing tasks performed in NLP applications.

2.3 Language Model Creation

A language model, put simply, is a statistical model that, given a set of words in a sequence, outputs the next words in the sequence. We use a method called **ULMFit (Universal Language Model Fine-Tuning)** [5], which involved using a pre-trained language model. Please see Section 3 for a detailed explanation of this language model. This model was originally trained on Wikipedia, and is hence able to learn a lot about how language is used in real life. We fine-tune the model on the collected Twitter data, and the model is able to learn about sexist language and how the short text that tweets are comprised of is used.

2.4 Classifier Creation and Usage

We create a classifier on top of the above language model to classify the labelled Twitter data as either sexist or not. We then use this classifier to classify our amateur fiction data. This is well-illustrated in the picture from the original ULMFit paper shown in Figure 1.

3 Background

In this study, we use NLP, or Natural Language Processing, techniques in order to map trends of sexism in modern amateur fiction sourced from social networking site Reddit, through a process called transfer learning, which is a paradigm in the field of machine learning, which in turn is, when defined informally, the process of making computers can learn patterns from large amounts of data. Transfer learning involves using a pretrained model to make inferences about data that the model has not been trained on. We provide some background below.

3.1 Natural Language Processing (NLP)

NLP is a sub field of computer science with a strong relationship to the field of artificial intelligence, and primarily refers to the process of using computers to analyze and generate text that simulates natural language usage in human beings. The field of NLP has a rich and varied history, with a variety of challenges being undertaken, such as those of neural machine translation, speech recognition, natural language generation, etc. In its early days, NLP tasks were usually achieved by using a variety of hand-coded rules, which involved encoding natural language syntax and semantics by hand by experts, an expensive, time-consuming and ultimately imperfect method, since it was prone to being biased towards the kind of language that the specific expert used, as well as his/her personal 'correct'

version of various colloquialisms. However, statistical methods that use large amounts of data to construct models simulating natural language soon found provenance over the hand-coded methods. These methods have been in use for a few decades now, but with the increasing amount of machine learning methods in play today, especially neural networks (see below), these methods have slowly begun to replace statistical methods for cutting-edge NLP research.

3.2 Machine Learning and Neural Networks

"A computer program is said to learn from experience E, with respect to some class of tasks T and performance measure P if its performance at tasks in T as measured by P improves with experience E. This definition of machine learning is perhaps the most common, and it simply means that a machine learning algorithm is a computer program that learns from data. With two major sub types, supervised (with labelled data) and unsupervised (with data that is unlabelled), the field of machine learning is used for a variety of tasks two of which are classification (placing objects into categories), and clustering (grouping similar objects together).

In our paper, we attempt to create a machine learning model to detect sexism in text. We use a modern machine learning algorithm called neural networks, which are loosely inspired by the human brain. A neural network is an algorithm that is loosely modelled on the human brain and has, in recent years, become the de facto tool in natural language processing research. A neural network involves layers of computational units called neurons, each layer connected to the next through weights and biases, and each neuron contains a nonlinear function called the activation function, and thanks to these non linearities, the neural network is found to be a universal function approximator [7], or put simply, any function can be approximated by a neural network with enough layers of neurons.

3.3 Classification and Transfer Learning

Classification is the process of placing things in categories. In machine learning, the process of classification, when defined mathematically, is that of creating boundaries between sets of data points that belong to the same category. It is found that neural networks are excellent at performing classification tasks, and are hence the tool of choice for our research work.

Transfer learning [8], on the other hand, is the process of building a model using a certain set of data for a task, and then using the model, with slight modifications, in order to perform different tasks. This may seem counter intuitive way of going about the process of creating a machine learning system, but it is an extremely powerful tool. Especially in cases in which we have large pre-trained models available to us, and we only need to slightly edit their capabilities, it would be a waste of both time and compute to create a new model from scratch.

4 Universal Language Model Fine-tuning

Inductive transfer or the ability of learning performance has been unsuccessful for NLP tasks. Universal Language Model Fine-tuning (ULMFiT) addresses the issues to teach a Language Model (LM) to adopt wider domain and enables robust inductive transfer learning for any NLP task, akin to fine-tuning ImageNet models. ULMFiT pretrains a LM on a large general-domain corpus and fine-tunes it on the target task using novel machine learning techniques. The model is universal in the sense that it meets:

- It works across multiple tasks and deals with varying document size, number and label
- Single architecture is used for training.

ULMFiT uses ASGD Weight-Dropped LSTM (AWD-LSTM) , a regular LSTM with no attention mechanism, with same number of hyper-parameters but added layers of Dropout. ULMFiT consists of following steps:

- **General-Domain Pretraining:** In this step the model is trained on very large and diverse dataset to capture the generality. Wikitext-103 dataset contains 8,595 preprocessed Wikipedia articles and 103 million words. This is computationally most expensive stage but needs to be done only once.

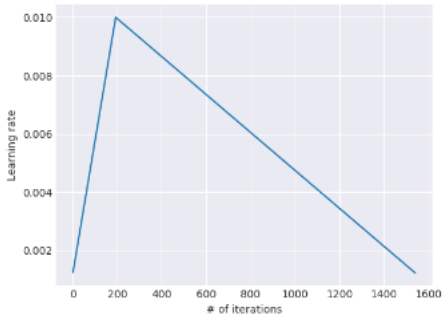

Figure 2: Slanted Triangular learning rates used for ULMFiT

- **Target task LM fine-tuning:** Although the model has learned diverse data domain but target task will come from different distribution. Thus LM is fine-tuned to learn the specific task. This stage converges faster and needs to adapt the data distribution. Two fine tuning methods:
  - **Discriminative Fine Tuning:** It is motivated by the fact that different layers holds different information. ULMFiT proposes that different layers to have different learning rate.
  - **Slanted triangular learning rates:** It refers to special learning rate scheduling where the learning rate increases linearly for a short time so that the model can converge to a parameter space suitable for the task fast and then learning rate decreases linearly for long time so that model fine tunes to the target task.
- **Target task classifier fine-tuning:** To fine-tune the model the architecture is augmented with two blocks. Each block uses batch normalization [14] and dropout, with ReLU activations for the intermediate layer and a softmax activation that outputs a probability distribution over target classes at the last layer.
  - **Concat pooling:** It extracts max-polling and mean-pooling over the history of hidden states and concatenates them with the final hidden state.
  - **Gradual unfreezing:** It helps in learning long-term dependencies in the language model by gradually unfreezing one layer at a time. First, the last layer is unfrozen and fine tuned and simultaneously the lower layers are unfrozen which helps in fine tuning each layer.

5 Future work

There are many directions in which the study of sexism in amateur literature can be taken. For one, our initial classifier can be trained using a corpus of literary documents, which would be much more indicative of amateur literature. The process of the creation of such a corpus would involve hand-labelling data as sexist or not, and could then perhaps be used to get a much better idea of sexism in long-form text.

6 References

1. `https://www.reddit.com/r/WritingPrompts/`
2. Samuel, Arthur (1959).*Some Studies in Machine Learning Using the Game of Checkers.*, IBM Journal of Research and Development.
3. `https://developer.twitter.com/en/docs/tweets/search/api-reference.html`
4. `https://www.tweepy.org`
5. Howard, Jeremy and Ruder, Sebastian (2018) *Universal Language Model Fine-tuning for Text Classification.*