TED vs Ted: Determining Professionalism of Text

Michael Arkhangelskiy, Carwyn Collinsworth, Michael McNeill

Problem Definition

The objective of our project is the binary classification of professional and unprofessional text. Professionalism is regarded as a subjective measure, but it is commonly thought to be related to syntactic complexity, corpus diversity, and other word-level measures.

Professional Text Input Example - “What we know about the brain is changing at a breathtaking pace”

Unprofessional Text Input Example - “I retired at 50 and I just turned 65, very happy to do so”

Motivation

- There are no existing objective measures of professional text. Most measures asses text with a paragraph or full-text scope, quantifying organization, flow, and coherence, however none measure it on a sentence level. Similar tasks use binary classification to discern authorship or feedback quality.
- This task is beneficial as the trained model can be used for verification on the quality of a piece of writing. Furthermore, a model to detect this could contain valuable insight into which attributes lend themselves to professional text.

Main Ideas

Our project worked to answer the following questions (see analysis for results):

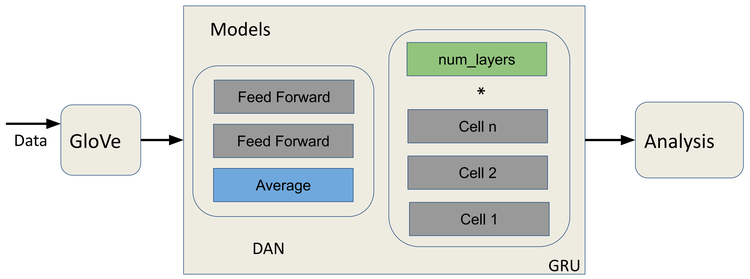
- Is it possible to train a neural classifier to decipher between professional and unprofessional text?
- Is “professional text” equivalent to syntactic complexity?

Sources

1. Elliott Casal, Joseph J. Lee. (2019) Syntactic complexity and writing quality in assessed first-year L2 writing, Journal of Second Language Writing, Volume 44, Pages 51-62, ISSN 1060-3743.
2. Lu, X. (2010) Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474-496.
3. Ötleş, Erkin et al. (2021) Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies. Academic Medicine 96: 1457 - 1460.
4. Noura Khalid Alhuqail (2021) Author Identification Based on NLP European Journal of Computer Science and Information Technology Vol.9, No.1, pp.1-26
5. Jarvis, Scott & Grant, Leslie & Bikowski, Dawn & Ferris, Dana. (2003) Exploring multiple profiles of highly rated learner composition. Journal of Second Language Writing. 12. 377-403. 10.1016/j.jslw.2003.09.001.

Method Details

- We implemented two different types of models: Deep Averaging Network (DAN), and Gated Recurrent Unit (GRU). Each of these models took in GloVe embeddings with an embedding dimension of 50, and had 4 layers.
- We also implemented an attention component within GRU to attempt to further improve the accuracy by focusing on important parts of the sentences.



Evaluation Setup

- We collected caption and comment transcripts from YouTube videos for our dataset. We operated under the assumption that comments are categorized as unprofessional text and the video transcript as professional. To help filter for higher quality data, we removed non-ascii characters from comments, and broke them down into sentences. This was not necessary for the video transcripts because they were well formatted. We choose to only include videos with closed captioning, and did not use any with automatically generated transcripts. We then used only sentences from the transcripts/comments which were longer than 5 words and shorter than 25 words.
- We used three separate datasets, each having the number of comment and transcript sentences equalized, to be half of the total for each run. We had 80% of the data used for training, 10% for validation, and 10% for testing. The datasets were TED talks (4104), lectures from Ivy league universities (5248), and news broadcasts (3490). We also trained and tested the model with a composite dataset by combining these smaller datasets.
- Our evaluation measures were: accuracy (percent correctly classified), precision (mislabel unprofessional text as professional), recall (ability to find all professional text), and F1 (combination of precision and recall).

Key Results & Analysis

TED Talks

Model	Accuracy	Precision	Recall	F1
GRU	0.7664	0.7183	0.8095	0.7612
GRU + Attention	0.6399	0.5928	0.6931	0.639
DAN	0.7445	0.6736	0.8624	0.7564

News Broadcasts

Model	Accuracy	Precision	Recall	F1
GRU	0.7765	0.7927	0.7471	0.7692
GRU + Attention	0.7307	0.7128	0.7701	0.7403
DAN	0.765	0.8538	0.6379	0.7303

Ivy Lectures

Model	Accuracy	Precision	Recall	F1
GRU	0.8403	0.8	0.9091	0.8511
GRU + Attention	0.8327	0.8492	0.8106	0.8295
DAN	0.8593	0.8467	0.8788	0.8625

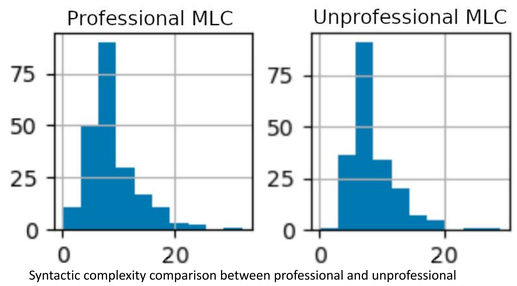
Composite

Model	Accuracy	Precision	Recall	F1
GRU	0.832	0.8217	0.8323	0.8269
GRU + Attention	0.8445	0.8523	0.8194	0.8355
DAN	0.8165	0.7927	0.8387	0.815

At low amounts of data, GRU and DAN perform similarly, with GRU + Attention lagging behind. As number of training samples increases, all models perform better. GRU grows faster than DAN, and GRU + Attention grows the fastest. With our best model on the composite dataset, GRU + Attention performs better than GRU and DAN, with a 0.8445 accuracy.

Our results align with related work. In [3], textual feedback is classified into helpful and unhelpful. With an SVM discriminator, they achieved 0.83 accuracy which is almost exactly the accuracy we were able to achieve.

When attempting to hand label among ourselves, we were only able to achieve 75.5% accuracy.



Syntactic complexity comparison between professional and unprofessional distributions. The measure used here is Measure of Lexical Complexity (MLC).

Lu [2] developed certain measures of lexical diversity and syntactic complexity. We used these measures to create sentence distributions between professional and unprofessional text. Differences between the distributions existed, but none were statistically significant. The figure above shows one of these measures, which shows a high similarity of syntactic complexity.

All 14 of the measures determined that the syntactic complexity distribution of professional text was indistinguishable from the distribution of unprofessional text. We conclude that our model is not learning how syntactically complex the text is. Therefore measuring “professionalism” of text is not equivalent to measuring syntactic complexity of text.

Misclassified Text: “But, I use the word skill because I believe it can be trained.” (Professional) “Additionally, their morale is very low and we saw that as well.” (Unprofessional)

Conclusion & Future Work

- We achieved comparable accuracy to baselines, and determined our models were not simply classifying based on syntactic complexity.
- We would like to expand our datasets to include other sources outside of YouTube, as this was operating under the assumption that all comments would be unprofessional, and transcripts would be professional, but even with a curated dataset such as ours this may not be guaranteed.
- We would also like to investigate which words are more heavily correlated with either professional or unprofessional.
- Improve cleaning of data to remove any remaining artifacts.