TED or Ted: Determining Professionalism of Text

Michael Arkhangelskiy
112119654
markhangelsk
@cs.stonybrook.edu

Carwyn Collinsworth
112605735
crcollinswor
@cs.stonybrook.edu

Michael McNeill
112338842
mmcneill
@cs.stonybrook.edu

1 Proposal Description

The objective of our project is the detection of professional vs unprofessional text. This could be beneficial as the trained model can be used for verification on the quality of a piece of writing. This problem is difficult to solve as there may be many different styles of writing, many of which are professional. Using machine learning we can find the overlap in these writing styles.

Background

Relevant research in this field consists of text quality classifiers, authorship detection classifiers, and other similar classifiers such as those for sentiment analysis. Below, we briefly summarize three works.

- [1] : Research concluded that there is a strong correlation between syntactic complexity [2] and writing quality. This finding and L2 Syntactic Complexity Analyzer are tools we can use to judge our models.
- [3] Using feedback collected from surgical residencies, a SVM model was built which had a 84% accuracy classifying between high and low quality feedback.
- [4] This article focused on authorship detection with different algorithms and feature extraction methods. Using bag of words with each of the algorithms tested (logistic regression, SVM, random forest, BERT) had the highest classification accuracy compared to latent semantic analysis.

In [5], a dataset of compositions from ESL placement students were clustered upon twenty-one linguistic features including relative frequency of amplifiers, hedges, impersonal pronouns, etc. It was determined that each of the five clusters (optimal) were separated mainly by differences in mean word length, nouns and nominalizations, prepositions, and present tense verbs (all relative frequency). Note that the clusters successfully discriminated between low, and medium-high quality compositions. The other features did not have statistically significant differences across multiple datasets, although some were useful in categorizing a single dataset. They also conclude that "the results of the present study suggest that the quality of a written text may depend less on the use of individual linguistic features than on how these features are used in tandem". The study focused on the importance of phenomenons called "complementary and compensation. Complementary refers to the fact that, although there may indeed be a number of linguistic features that contribute to the overall quality of a written text, high levels of some features may bring about low levels of other features". Similarly, compensation "refers to the idea that successful writers may be able to compensate for potential deficiencies in their writing by capitalizing on a few of their strengths". Overall, they surmised that individual features cannot be indicative of quality of text, and noted that text quality is almost certainly biased upon the overall length of the text. In our study, we overcome this bias by constraining the length of our text segments.

Ideas/methods to address [Not included in proposal]

Testing

We can use a variety of training and test sets. One plan would be to take the transcript and comments section from a TED talk. The speaker would provide the professional data, and the comments could provide the unprofessional data which we would use for training our model. Since these data sets come as a pair, we know that they

will still cover the same topic material, removing that bias. This method would generalize to other sets of YouTube videos, allowing for a more diverse data set on different topics. To obtain more information for training, we will filter out shorter length comments, as well as splitting the transcript into multiple parts. To ensure the transcript is broken in a cohesive way, we will only use videos which have manual subtitles, and not use any videos with auto-generated captions.

Evaluation

In order to determine the effectiveness of our model and determine if it is learning what we want it to, we will use the following tasks:

- **Probing:** We will change parts of a normally (un)professional text and examine how changing words or chunks of the sentence affects the end classification. Specifically, using data from yet-unseen sources will be used to test the models.
- **Compare sentence improvement:** We will start by taking groups of informal sentences, then using a tool such as Grammarly, we will obtain a more professional sentence. We can then compare the improved sentence to see if the classification changes.
- **L2 Syntactic Complexity Analyzer:** This python package can be used as a benchmark for evaluating the prediction power of our models. While syntactic complexity is not the direct measurement our models predict, they should still recognize this as a quality of professional vs unprofessional writing.

Research Questions

- Results Table: We will compare our results to the SVM and bag of words results from [4].

Model	Feature	Accuracy	Precision	Recall	F1
GRU	GloVe				
GRU + attention	GloVe				
DAN	GloVe				
SVM	Bag of Words				

- Research question 1: Does syntactic complexity correlate with the professionalism of text? Using the L2 Syntactic Complexity Analyzer provided by [1], we can compute the syntactic complexity of professional and unprofessional text from our test dataset. We can then compute a measure of correlation between our own classifier and this data, such as spearman correlation coefficient to draw conclusions about similarity of classification.
- Research question 2: How does sequential representation of words (RNN / GRU) differ from non-positional aggregations of words (DAN).
- Research question 3: Does attention improve our models classification ability? This will be tested by comparing classification results between GRU with an attention component and GRU without an attention component.
- Research question 4: Does the average length of words in a sentence correlate with the professionalism of text?

Goals [Not included in proposal]

Phase I Progress

1. **System Details:** We have modified our second homework to have the ability to use multiple attention heads. We will utilize multiple attention heads for the GRU and use DAN as our baseline. These are trained on a set of `.jsonl` files containing sets of texts with the labels of 1 (transcript) and 0 (comment). These files will then be used for training, validation, and testing. Our data is collected using YouTube comment and transcript scrapers, which then break these into individual sentences. We prune out sentences which are too short, and remove non ASCII characters such as emojis.
2. **Implementation Progress** We have been able to collect a small set of data and train a few small models, but have only been able to get accuracy ranging from 65-75% on different runs with different data split sizes. Our most recent models have testing accuracies of 75% (DAN), 73% (GRU), and 74%(GRU with 10 attention heads).
3. **Analysis** The model appears to work mostly as expected, although it has lower accuracy than we had originally imagined when we started this project. However, manually classifying our dataset we had an average accuracy of around 60%. One issue we currently

have is that adding attention to the GRU sometimes decreases accuracy. For different splits of training, validation, and testing GRU with attention can outperform DAN, but it appeared inconsistent. This may change once we tune our hyperparameters.

4. **Questions/Concerns** The plan still sounds viable, but we had a number of different ideas compared to originally. We believe that the task may be too difficult to classify using only one sentence. We plan to run more tests using 2-3 sentences grouped together and see if the models perform better. We also want to include content outside of TED talks for our professional datasets (news broadcasts, political talks). TED talks have a certain structure and we want to vary our data.

References

1. J. 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.