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.

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				
	Bag of Words				
SVM	GloVe				
	Bag of Words				
SVM (Baseline)	Bag of Words				

- Research question 1: Can varying feature extraction help improve performance? In our research, we will vary the feature extraction methods in effort to classify with the highest accuracy. We have chosen Bag of Words and GloVe as our feature extraction methods.
- Research question 2: Can varying machine learning model type help improve performance? In our research, we will vary the model types in effort to classify with the highest accuracy. We have chosen SVM and GRU as our machine learning models.

Goals [Not included in proposal]

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