Linguistic Twins: Utilizing Siamese Networks for Effective Author Style Discrimination

Cristina Velázquez
Facultad de ingeniería
Universidad Panamericana
Aguascalientes, México
0224433@up.edu.mx

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Analyzing an author's writing style refers to identifying the frequency with which style markers are used. Some examples of these markers are frequent sequences of words, typing errors, and punctuation marks. This type of analysis focuses on how an author uses language features. It can be either extrinsic where a comparison between authors is made or intrinsic in which the author's writing style is compared through different stages of their life.

Identifying the author of any given text can lead to multiple applications such as plagiarism detection, book recommendation systems, forensic applications, and even fake news detection.

Right now there are multiple fantasy authors and even more people who dabble in writing fanfiction related to these fantasy books. Generally, fantasy novels tend to treat similar topics, scenarios, places, and even magic systems. This means that in most fantasy novels many words are frequently used, no matter who the author is.

These conditions create an interesting standing point when trying to determine if two excerpts of fantasy novels were written by the same person, as it is necessary to understand and identify the differences in the writing style of each author. Correctly detecting the differences in fantasy authors' writing styles can lead to many applications. As was mentioned before, fantasy novels tend to be similar to one another, so this means that the factor that determines whether an author reaches worldwide success or not is entirely dependent on the way each author narrates stories. On the writing style each author has.

Later on, a new investigation can be made, this one focusing on detecting what aspects are common between famous's authors' writing styles. By finding these characteristics, it would be possible to teach new authors how to properly write. This gives them a bigger chance of reaching worldwide success with their stories, all because of how they narrate those new stories. A new generation of authors can be born out of this type of investigation.

This is why, in this paper, the main focus is testing the performance of a proposed architecture when detecting if two different samples of fantasy fanfiction text were authored by the same person or not.

The dataset used is made out of fantasy fanfictions as this allows for a deeper investigation. When a novel is published it must go through a series of revisions, all to avoid typing errors, unclear sentences, or unnecessary punctuation marks. Since these things are style markers that help to identify an author's writing style a good starting point is to use text fragments published or shared by people without the added process of editing. Because of this, the dataset used is one put together by Fernandes et al. [1]. It was obtained from fanfiction.net, gathering text fragments made by 10,000 different authors, all of them belonging to one of the most popular fandoms the platform contains, which are: "The Hunger Games", "Percy Jackson and the Olympians", "Lord of the Rings", "Harry Potter", and "Twilight".

In most cases, there is only one text sample per author and the samples contain around 2,000 words each. it's important to notice that since these texts are obtained from a fanfiction platform, they tend to contain many typing errors, which will be helpful when learning the author's writing styles. Because all of the samples belong to certain fandoms, it is ensured that the general topic tends to be similar, thus, many of the same words are frequently used. This will be a challenge, as frequent words are also something to consider when studying a writing style.

Yet, the purpose of this paper is to understand how the proposed architecture performs given these conditions.

For the authorship attribution problem, multiple approaches have been proposed and can be divided into two sections, classification-based and similarity-based. In this paper, the focus will be the similarity-based approach, as a metric is used to measure the distance between two texts, and it's possible to recognize whether they were written by the same author or not. The great advantage of this method is that it allows for a more generalized solution, as it does not need a large number of texts authored by a certain person to be able to recognize if a new text sample not present on the training set belongs to them.

To define the proposed architecture in this paper, previous works were revised, the most relevant ones, due to the similarity they share on the problem to solve are the following. Mueller and Thyagarajan [3] used a siamese adaptation of the LSTM network to assess the semantic similarity between sentences. The results were promising as they demonstrated that LSTMs are powerful language models capable of tasks that require intricate understanding and they preceded the state of the art on this problem. Inspired by this previous work, C. Saedi and M. Dras [4] used a siamese network to produce similarity scores for text pairs such that pairs by the same author have high scores and those by different authors have lower scores. They use CNNs and LSTMs as subnetworks and demonstrated that the proposed architecture performs well when solving the previously mentioned problem. Finally, Sukanya Nath proposed a siamese neural network for learning the stylistic similarity between two short texts (mean size of 50 words). The results were positive and demonstrated that style change detection may be cast as a one-shot learning task. More details of each mentioned work will be presented in the state-of-the-art section.

A SNN comprises two identical twin networks with the same weights that learn the hidden representation of two different input vectors. The similarity of the output of both these networks create is calculated using a distance measure. This paper's end goal is for a SNN to produce similarity scores for text pairs such that pairs authored by the same person are given high scores and those that are written by different people obtain lower scores.

The subnetworks proposed are transformers instead of LSTMs or CNNs, also there are changes from the previously mentioned works on the size of the texts used as input for the network, the data preprocessing, and finally, the performance of various distance functions is compared.

II. State of the art

A. Siamese neural networks

Bromley, J., et al [7] use siamese networks for signature verification, by making it an image-matching problem. For general one-shot image recognition, Koch et al [8] used convolutional siamese neural networks. A. Abdalhaleem et al [11] presented an automatic system for dividing a manuscript into similar parts, according to their similarity in writing style. In the training, the two sub-networks extract features from two patches, while the joining neuron measures the distance between the two feature vectors. Patches from the same page are considered identical and patches from different books are considered as different. Based on that, the Siamese network computes the distances between patches of the same book. From works like these, inspiration came and the use of SNN was adapted for use in NLP tasks.

B. Siamese neural networks for NLP tasks

The works mentioned in this subsection are those that inspired this's paper proposed architecture.

Mueller and Thyagarajan [3] used a siamese adaptation of the LSTM network and they show that given enough data, a simple adaptation of the LSTM may be trained on paired examples to learn a highly structured space of sentence representations that captures rich semantics. Their model uses an LSTM to read word vectors representing each input sentence and employs its final hidden state as a vector representation for each sentence. Subsequently, the similarity between these representations is used as a predictor of semantic similarity. Finally, to measure the similarity between representations they use the Manhattan distance. Inspired by this previous work, C. Saedi and M. Dras [4] presented an investigation of the application of siamese network architecture to large-scale stylistic author attribution. The goal was to produce similarity scores for text pairs such that pairs by the same author have high scores and those by different authors have lower scores. They compared the performance of the model when using CNNs and LSTMs for the sub-networks, and also examine the effect of using world-level input or character-level input. For measuring the distances between the texts, cosine similarity and L1 functions are used. Sukanya Nath [2] proposed a siamese neural network for learning the stylistic similarity between two short texts (mean size of 50 words). The sub-networks used are bidirectional LSTMs and GRUs. In this case, the cosine distance function was proposed for measuring the distance between the texts. Overall the results are positive in all the previous works, yet there is mention of the long time LSTMs take to train.

This paper's proposed architecture uses transformers as the sub-networks instead of LSTMs, CNNs or GRUs to compare the performance obtained when comparing two text samples. The goal is to reduce the training time. As for the distance functions, cosine similarity and L1 are both used. The modifications on the dataset put together by Fernandes et al. [1] and the preprocessing used are explained in the following section.

III. Dataset

The dataset used is made of text fragments made by 10,000 different authors, all of them belonging to one of the most popular fandoms the platform contains, which are: "The Hunger Games", "Percy Jackson and the Olympians", "Lord of the Rings", "Harry Potter", and "Twilight". In most cases, there is only one text sample per author and the samples contain around 2,000 words each.

The data is stored in 10,000 folders, each folder containing one text sample of a given author, whose name is specified by the folder's name. Pairs of text samples must be created and labeled. The label is 1 in case both texts are written by the same person and 0 in case they are from different authors. To achieve this, the following steps are followed:

- A random number is created to select which author will be used to obtain the first text sample.
- A parameter of 0.6 is established to determine the probability of selecting a file pair from the same folder. A random number between 0 and 1 is obtained. If the random-number is smaller than the probability both samples are obtained from the same author and y = 1. Otherwise, two different authors are used, and y= 0.
- In case y=0 another random number is created to select the second author from whom the second text sample will be obtained.
- When the text files are selected, they are read, turned into lowercase, and split into words.
- To try and use a data augmentation technique, a random point is selected for the split text to create a sequence of words of a maximum length of 128. This is to obtain different text samples each time an author is selected. Then the words are encoded using the BERT tokenizer, with special tokens added to mark the start and end of the sequence.
- The method returns a tuple of encoded input sequences and a similarity label y. The two input sequences correspond to two text files selected according to the probability parameter and the similarity label is set to 1 if the two files come from the same folder (author) and 0 otherwise.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] Fernandes, N., Dras, M., McIver, A., 2019. Generalised differential privacy for text document processing. In: Proceedings of Principles of Security and Trust (POST), LNCS, Vol. 11426, pp. 123–148.

[2] Nath, Sukanya. (2021). Style change detection using Siamese neural networks (Notebook for PAN at CLEF 2021).

[3] J. Mueller, A. Thyagarajan, Siamese recurrent architectures for learning sentence similarity, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[4] Chakaveh Saedi, Mark Dras, Siamese networks for large-scale author identification, Computer Speech Language,Volume 70,2021,101241,ISSN 0885-2308,https://doi.org/10.1016/j.csl.2021.101241.

[5] N. Schaetti, Character-based convolutional neural network and resnet18 for twitter authorprofiling, in: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France, 2018, pp. 10–14.

[6] A. Iyer, S. Vosoughi, Style change detection using bert, in: CLEF, 2020.

[7] Bromley, J., Guyon, I., LeCun, Y., Sackinger, E., Shah, R., 1994. Signature verification using a “siamese” time delay neural network. Advances in Neural Information Processing Systems, pp. 737–744.

[8] Koch, G., Zemel, R., Salakhutdinov, R., 2015. Siamese neural networks for one-shot image recognition. ICML Deep Learning Workshop, Vol. 2.

[9] Rios-Toledo, G., Posadas-Duran, J. P. F., Sidorov, G., Castro-Sanchez, N. A. (2022). Detection of changes in literary writing style using N-grams as style markers and supervised machine learning. PLoS ONE, 17(7 July), [e0267590]. https://doi.org/10.1371/journal.pone.0267590

[10] Lorenzen, S. S., Hjuler, N. O. D., Alstrup, S. (2019). Investigating Writing Style Development in High School. In Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019) (pp. 572-575). Université du Québec à Montréal.

[11] A. Abdalhaleem, B. K. Barakat and J. El-Sana, ”Case Study: Fine Writing Style Classification Using Siamese Neural Network,” 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), London, UK, 2018, pp. 62-66, doi: 10.1109/ASAR.2018.8480212.