Replikácia článku: Klasifikácia suicidálnych príspevkov pomocou CNN

Porokh, Uhrynchuk
Technical University of Košice

Abstract

Tento dokument predstavuje replikáciu článku *Analyzing Social Media Texts for Suicidal Risk Identification using Natural Language Processing*. Cieľom je vytvoriť nezávislú implementáciu CNN modelu na základe popisu v článku a porovnať výsledky s pôvodnou prácou. Na spracovanie textových údajov sme použili NLP techniky a predtrénované vektory GloVe. Hodnotíme presnosť modelu a diskutujeme o možných vylepšeniach.

I. Úvod

Suicidálne sklony predstavujú významný verejno-zdravotný problém. Sociálne siete poskytujú cenný zdroj informácií o mentálnom stave jednotlivcov. Cieľom tejto práce je použiť NLP a strojové učenie na klasifikáciu textov s cieľom identifikovať riziko samovraždy.

II. Zhrnutie pôvodného článku

Pôvodný článok prezentuje použitie viacerých klasifikačných modelov (SVM, Logistic Regression, CNN) na analýzu textových príspevkov zo subredditu *SuicideWatch*. Príspevky boli reprezentované pomocou TF-IDF a GloVe embeddingov.

Hlavná pozornosť bola venovaná CNN architektúre, ktorá obsahovala:

- Vstupnú embedding vrstvu,
- 1D konvolučnú vrstvu s 128 filtrami a veľkosťou kernelu 5,
- GlobalMaxPooling vrstvu,
- a výstupnú Dense vrstvu so sigmoid aktivačnou funkciou.

Model dosiahol presnosť 92.09%, pričom podľa autorov prekonal ostatné prístupy. Dataset však nebol verejne dostupný, čo sťažuje presnú replikáciu.

III. Naša implementácia

A. Spracovanie údajov

Na čistenie textov sme použili knižnice `spaCy` a `NLTK`, kde boli aplikované kroky ako:

- Konverzia textu na malé písmená,
- Odstránenie interpunkcie a špeciálnych znakov,
- Tokenizácia a odstránenie stop-slov,
- Lemmatizácia slov.

Kód pre spracovanie je uvedený v súbore `data_preprocess.py`.

B. Embeddingy a tokenizácia

Použili sme GloVe embeddingy (100 dimenzionálne, `glove.6B.100d.txt`) a `Tokenizer` s maximálne 10k slovami. Texty boli transformované na `padded_sequences` s dĺžkou 200. Bolo vytvorené embedding_matrix, ktoré bolo nahraté do embedding vrstvy modelu.

C. Architektúra modelu

Model definovaný v `model.py` obsahuje:

- Embedding layer (váhy z GloVe, nemenné),
- Conv1D (128 filtrov, kernel 5, ReLU),
- GlobalMaxPooling1D,
- Dense(1, sigmoid)

Tréning prebiehal s použitím optimalizátora Adam, BinaryCrossentropy ako loss funkcie, a metrikami: accuracy, precision, recall. Aplikovali sme aj `EarlyStopping` s cieľom zabrániť pretrénovaniu.

TABLE I
POROVNANIE CNN MODELU: NÁŠ DATASET VS. PÔVODNÝ ČLÁNOK

Metrika	**Náš model**	**Pôvodný článok**
Presnosť (Accuracy)	93.21%	92.09%
Precision	95.22%	91.65%
Recall	91.07%	92.56%

IV. VÝSLEDKY A POROVNANIE

A. Naše metriky

B. Vplyv EarlyStopping

Zaznamenali sme výrazné zlepšenie stability tréningu po použití EarlyStopping.

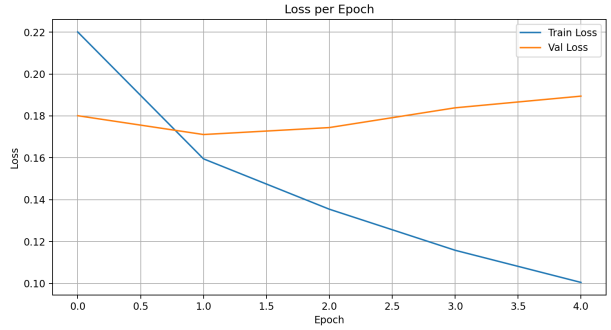

Fig. 1. Strata pred aplikáciou EarlyStopping

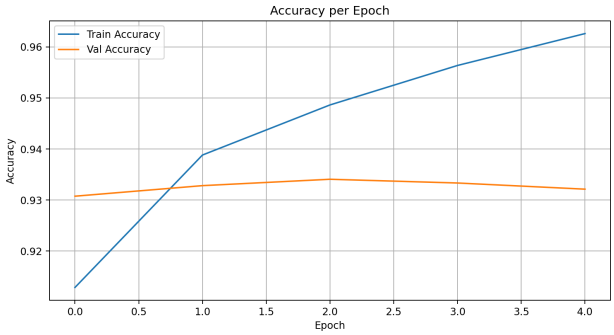

Fig. 2. Presnosť pred aplikáciou EarlyStopping

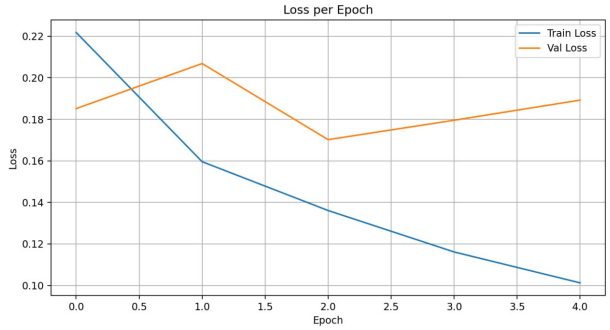

Fig. 3. Strata po aplikácii EarlyStopping

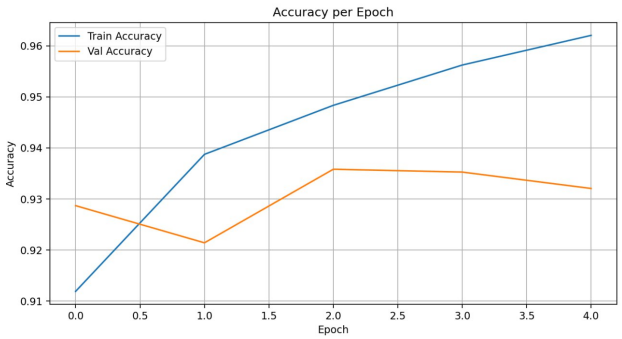

Fig. 4. Presnosť po aplikácii EarlyStopping

V. Rozdiely a odchýlky

- Pôvodný článok nešpecifikoval niektoré hyperparametre (napr. batch size, počet epoch), preto sme ich zvolili na základe best-practices (batch=64, epochs=5).
- Použili sme explicitné spracovanie textov pomocou moderných knižníc NLP.

VI. Kritické zhodnotenie

CNN model preukázal výborný výkon pri klasifikácii suicidálnych príspevkov. Avšak:

- Kontextové porozumenie by mohli zlepšiť modely typu Transformer (napr. BERT, RoBERTa).
- Možnosť použitia trénovateľných embeddingov pre lepšie adaptovanie na špecifiká dát.
- Pridanie Dropout vrstvy by mohlo zlepšiť regularizáciu.

VII. Záver

Replikácia potvrdila účinnosť CNN modelu v identifikácii suicidálnych príspevkov. Napriek rozdielom v datasete a architektúre sme dosiahli mierne lepšie výsledky než pôvodný článok. Práca demonštruje silu jednoduchých CNN architektúr v NLP úlohách.

References

[1] Shanskar Rai et al., “Analyzing Social Media Texts for Suicidal Risk Identification using NLP,” in *ICSC 2023*, IEEE, DOI: 10.1109/ICSC60394.2023.10441398.