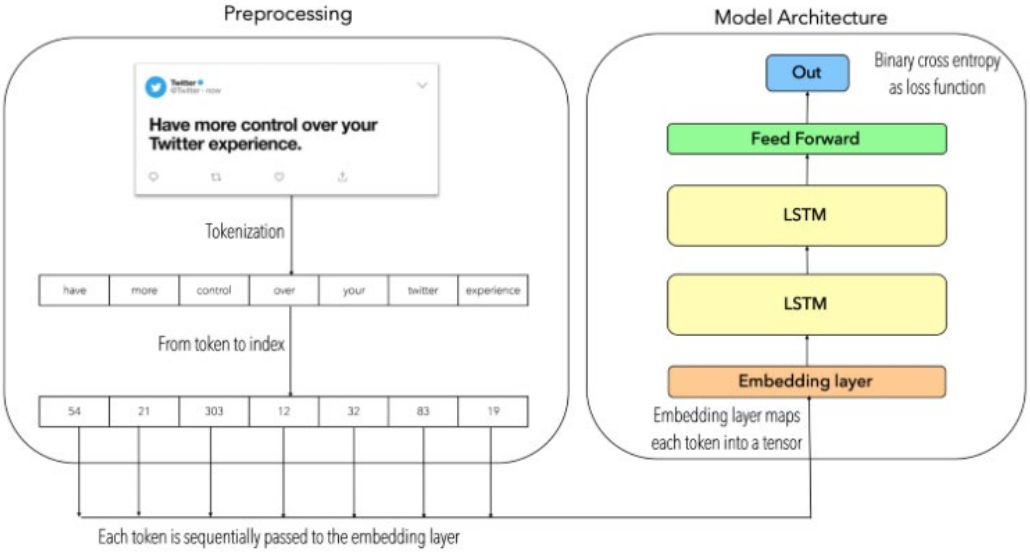
עבודת הגשה 3

Text Classification with LSTMs in PyTorch



לינק ל-GitHub:

מגישים:

מיכאל איפרגן 203842125

רון וקנין 305769440

סיווג טקסט

בעבודה השתמשנו בדאטה של ציוצים מטוויטר חלקם אמיתיים וחלקם לא. ראשית בתהליך הכנת הדאטה האלגוריתם משתמש בשיטת טוקניזציה שמפרקת כל ציוץ בדאטה לרצף של טוקנים, ובכך מכינה את הדאטה למודל. לאחר מכן יצירת המודל שיודע להעביר רצף טוקנים כזה לרשת נוירונים. כעת נותר לאמן את המודל, החלוקה בין הדאטה לאימונים ובדיקות הוא 50% לכל אחת מהקבוצות. בחרנו לבצע 15 סבבי אימונים שמדפיסים את פונקציית ההפסד, והדיוק של המודל עבור כל קבוצת דאטה. לבסוף הצגנו שני גרפים שמטרתם להראות את הירידה בפונקציית ההפסד בכל סיבוב אימונים וההבדלים בדיוק בין שני קבוצות הדאטה. ניתן לראות כי בעוד שדיוק קבוצת האימון עולה, דיוק קבוצת הבדיקה נשארת חסומה באזור ה-65-70%.

מסקנות והצעות לשיפור

השימוש בשיטת הטוקניזציה כדי לקחת ציוץ ולהפוך אותו לרצף טוקנים נפרדים שעוברים לאחר מכן לרשת נוירונים דו-שכבתית מראה בעצם את התהליך שהמידע עובר מאיך שהוא בקובץ לאיך שהמודל מקבל אותו.

ההצעה לשיפור שלנו היא לפצל את הדאטה לרמות קושי בזיהוי כלומר, ראשית לעשות סבבי אימונים עבור דאטה שיהיה למודל קל לזהות מה אמיתי ומה לא, לאחר מכן רמת קושי בינונית ולאחר מכן לאמן אותו ברמת קושי קשה.