מסמך מסכם – מיני-פרויקט בנושאים במדעי הרוח הדיגיטליים

דניאל סמוטריצקי, אורי סוכי

רקע

בקריאת טקסטים ספרותיים מכלל הז'אנרים, הסופרים והתקופות, ניתן להבחין שבכל טקסט יש את הנושאים המרכזיים שסביבם הטקסט נכתב ויש מאפיינים ש"מקשטים" את הנושא ותורמים ל"אווירה".

הפרויקט שלנו מתבסס על "סיפורים קצרים" ומטרתו היא לזהות ולסמן את המשפטים שפחות קשורים לגרעין של הסיפור אלא יותר מתארים את הסובב אותו.

לצורך הפרויקט הגדרנו את המושג "משפט תיאורי" כמשפט המתאר נוף, בגדים, מיקום, מסלול וכו'.

מטרת הפרויקט

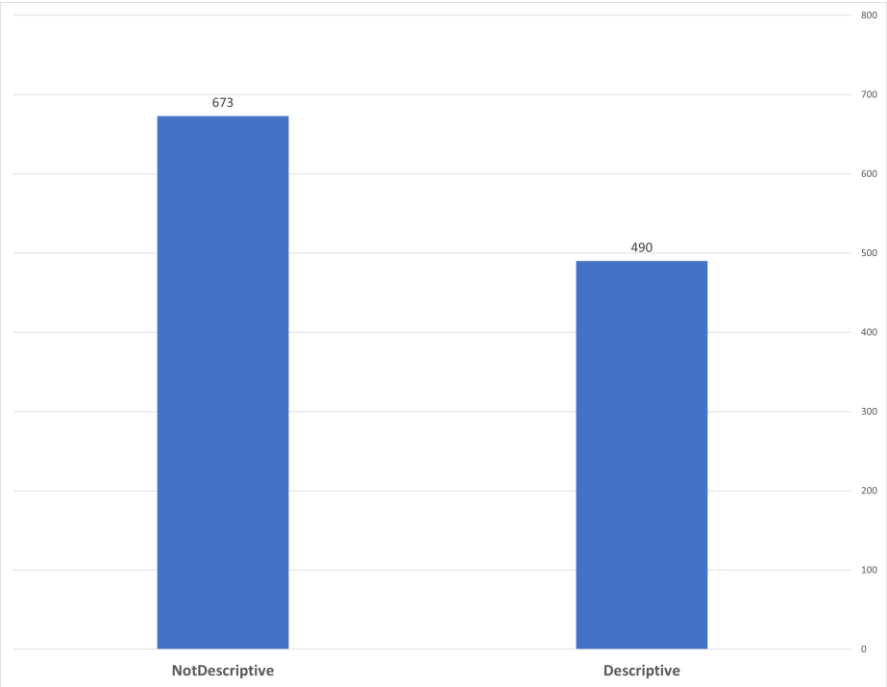
יצירת מודל לסיווג משפטים בטקסט כ"משפטים תיאוריים"

שיטה:

1. כתיבת קווים מנחים לתיוג משפט כמשפט "תיאורי"
לאחר קריאה של מספר סיפורים קצרים ניסחנו קווים מנחים להגדרת משפט כמשפט תיאורי.
   - משפט שלם המתאר נוף
   - משפט שלם המתאר חפץ / אובייקט
   - משפט שלם המתאר מסלול
   - משפט שלם המתאר מבנה

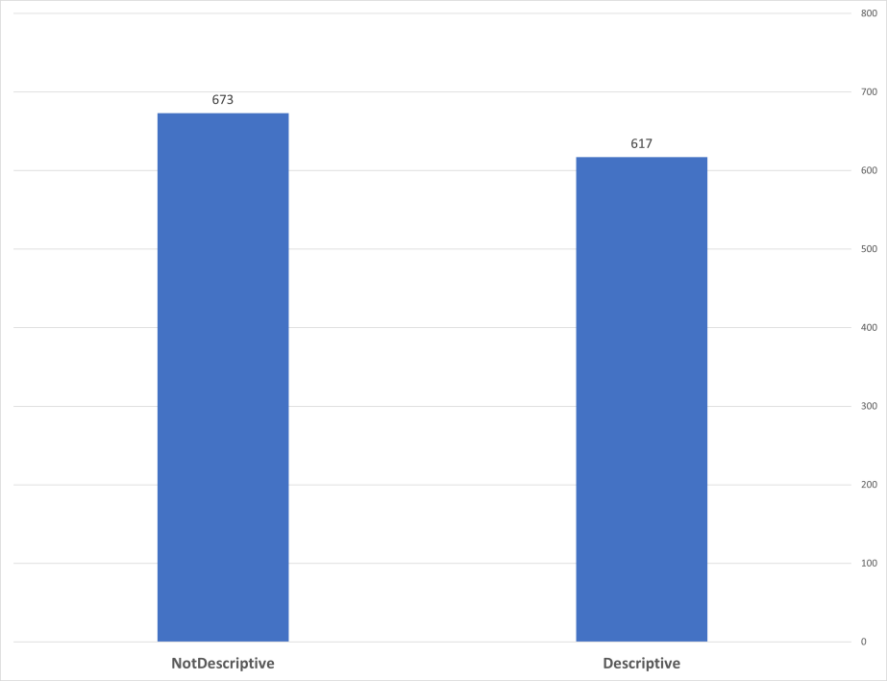
   קווים מנחים אלו נועדו לתיאום בינינו כמתייגים וכן ככלי עזר למי שירצו להבין ואולי לפתח את ה-dataset בעתיד.

2. תיוג ע"י קריאה מקרוב של מספר סיפורים קצרים מתקופות שונות
על מנת ליצור dataset מייצג כמה שניתן ניסינו, כמיטב יכולתנו, לאסוף סיפורים קצרים ממספר סופרים שונים ותקופות שונות.
3. יצירת dataset מהתיוגים
לאחר בחינת ה-dataset המבוסס רק על תיוג הסיפורים הבנו שה-dataset לא מספיק מאוזן:

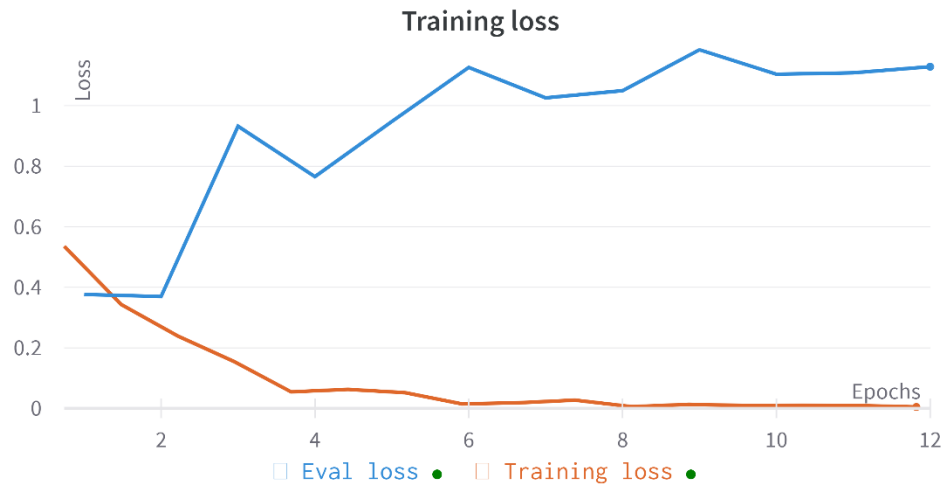


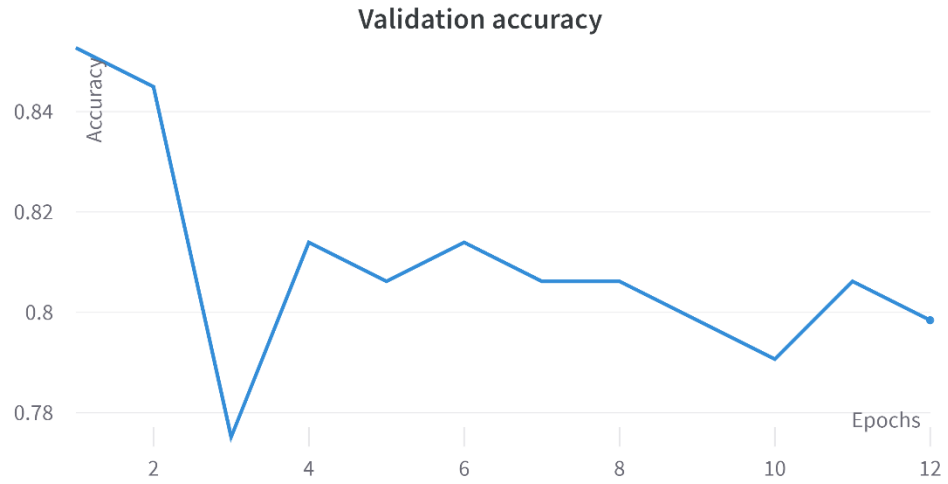
לכן החלטנו, לצורך האיזון, ליצור משפטים תיאוריים באופן מלאכותי.

לצורך כך השתמשנו ב-dataset הקיים, דגמנו מתוכו את המשפטים התיאוריים, בעזרת ה-API של DICTA ביצענו ניתוח מורפולוגי על המשפטים. זיהינו את המילים שהן או Nouns או Adjectives במשפט והחלפנו אותן ב-$[MASK]$.
בעזרת ה-pipeline של HuggingFace השתמשנו במודל Fill-Mask של AlephBert אותו הפעלנו על המשפטים וקיבלנו כפלט רשימה של חלופות אפשריות ל-$[MASK]$ שלנו.
בחרנו עבור כל משפט את מספר חלופות שהסבירות שלהן היא הגבוהה ביותר ויצרנו מאגר משפטים תיאוריים שנוצרו באופן מלאכותי. בשלב זה נדרש מעט ניקוי, בוצע באופן ידני. בחרנו מתוך המאגר כ-130 משפטים והוספנו ל-dataset שלנו.
לאחר איזון:



4. יצירת מודל סיווג טקסט על בסיס ה-dataset
לצורך בניית המודל בצורה יעילה ובמשאבים שיש לנו השתמשנו במודל מאומן AlphaBert בעברית.
ביצענו fine-tuning למודל ואימנו אותו על 80% מה-dataset שלנו. לצורך Validation השתמשנו ב-10% מה-dataset. התוצאות היו:





ניתן לראות שאומנם ה-Loss של ה-Validation הולך וגדל אבל ה-Accuracy יחסית גבוה.

5. בחינת המודל
בחנו את המודל על 10% מה-dataset שלנו והתוצאה הייתה כ-55%
6. פרסום ה-dataset והמודל
העלינו גם את ה-dataset המתוייג וגם את המודל לאתר Huggingface לשימוש הכלל:

- Descriptive_Sentences_He – Dataset
- Descriptive_Classifier - Model
- קבצים וקוד - GitHub

כלים:

1. תיוג:
   - Catma

2. מודלים לניתוח שפה
   - HuggingFace
   - AlephBERT
   - Dicta

מטרות ושימושים אפשריים:

- בחינה האם קיים קשר באופן ובתדירות השימוש במשפטים תיאוריים בין סיפורים שונים **באותו הז'אנר**
- בחינה האם קיים קשר באופן ובתדירות השימוש במשפטים תיאוריים בין סיפורים שונים **באותה התקופה**
- בחינה האם קיים קשר באופן ובתדירות השימוש במשפטים תיאוריים בין סיפורים שונים של **אותו הסופר**
  - ומתוך כך, מודל נוסף להבנת "טביעת האצבע" של סופר
- בחינה האם קיים קשר בין הופעת שם במשפט לבין סוג המשפט
- בחינה האם קיים קשר בין תדירות משפטים תיאוריים באזור מסוים בטקסט לתדירות מילות תואר באותו אזור
- עזר לסיכום של טקסט (Text Summarization) – לרוב משפטים תיאוריים בהכרח נדרשים פחות בסיכום.
- ועוד...

מסקנות:

בפרויקט זה יצרנו מודל לסיווג משפטים כמשפט תיאורי. מהתוצאות נראה שבהחלט ניתן בעזרת מודלים חישוביים להבחין בין משפטים תיאוריים וכאלה שלא.

לדעתנו, עוד ניתן לשפר בצורה משמעותית את התוצאה:

- לדייק את ה-guidelines ל-dataset ולשפר את התיוגים שלנו
- להגדיל את ה-dataset ולהוסיף מבחר יותר רחב של משפטים
- למצוא Hyperparameters טובים יותר למשימה

לשם כך העלינו את ה-dataset ואת המודל שלנו ל-Huggingface, כמו גם את כל תהליך העבודה שלנו ל-GitHub. מתוך תקווה שיהיה מי שירים את הכפפה וימשיך את העבודה שהתחלנו.