Whose Headline Is It Anyway?

IML Hackathon 2017

Ady Kaiser
Gilad Lumbroso
Omer Alon
Or Dagan

תהליך העבודה

1. <u>סריקת המידע</u>:
תחילה, הסתכלנו על כל הכותרות וניסינו להבין מה מבדיל אותן ואילו features נוכל להוציא מכל כותרת כך שנוכל ללמוד אותה. הדברים הראשונים שבלטו לנו הם - אורך הכותרות, ומילים שונות (למשל:) אשר מופיעות יותר אצל עיתון אחד מאשר השני.

2. <u>ניקוי המידע</u>:
שמנו לב כי למשל הכותרת "Haaretz Cartoon" חוזרת על עצמה פעמים רבות. כותרות כמו אלה לא מועילות לזיהוי תבנית הכתיבה של עיתון הארץ ורק יגרמו ל-over-fitting. לכן פילטרנו את המידע מכותרות בעייתיות.

3. <u>בחירת מודל הלמידה</u>:
החלטנו לעבוד במודל ה-stacking - אימנו מספר לומדים שונים בתחומים שונים ובשיטות למידה שונות. בסוף המודל הראשי שלנו יקבל את החיזוי של כל אחד מהם והחיזויים האלה יהיו מרחב הדוגמאות שאותו המודל ינסה ללמוד.

4. <u>בחירת המודלים הלומדים במסגרת ה-stacking:</u>
ניסינו למצוא מודלים שונים שישמשו כ-classifires בשלב הבסיס במודל ה-stacking. בחרנו מספר תחומי למידה, כאשר בכל תחום ניסינו מספר מודלים עם פרמטרים שונים. לבסוף בחרנו את התחומים והמודלים שהחזירו לנו את התוצאות הטובות ביותר. כל מודל בשלב הבסיס של ה-stacking נבנה כמחלקה נפרדת, באופן הבא:

   a. *Linear SVC - Bag of words*
   החלטנו למפות את המילים שנמצאות בכל כותרת ע"י וקטור שמייצג את המילים ומונה שסופר כמה פעמים כל מילה מופיעה. השתמשנו בטווח (1, 2) עבור פרמטר ה-ngram כדי ללמוד לפי מילים בודדות וצמדים.

   b. *SVC - Sentiment of text with the help of Google*
   בעזרת שירות "Google cloud Language API" שמנתח טקסטים, הוצאנו לכל כותרת את ציון ה-"Sentiment" שלה. כלומר, מספר בקטע [1, 1-] שמציין את עוצמת וחיוביות הרגש שהיא מבטאת. ציון של 1- הוא רגש שלילי ביותר, ציון של 1+ הוא רגש חיובי ביותר. בנוסף וקטור הנתונים מכיל את עוצמת הרגש במשפט בתחום [1, 0].
   רצינו להביע בחיזוי שלנו את הקשר בין ה"סטיגמות" על רצינות וביקורתיות "הארץ" לעומת

אהדה ושביעות רצון שמתבטאות בכותרות "ישראל היום". חשבנו שזה יהיה פיצ'ר שונה ומעניין לעומת פיצ'רים אחרים שמסתכלים בעיקר על מילים ולא על משפטים.
נציין כי שימוש בפיצ'ר זה, שדרש ניתוח טקסט בצורה טובה, לקח משמעותית יותר זמן מהאחרים. כיוון שלא הספקנו ונגמר לנו הזמן, אנו מגישים את הקוד של מחלקה זו כפי שהוא, כלומר שאינו עובד ביעילות ובמהירות וזאת כיוון שלא הספקנו לעבור על כל המאפיינים השונים במנתח הטקסט של גוגל ולכן לא יכולנו ליעל את הקוד. רצינו עדיין להעביר את הרעיון כי יש חשיבות לאווירה ולטון בטקסט ובניגוד הקיים בין שני העיתונים. נציין כי כעת, במתכונת הנוכחית שלוקחת הרבה זמן, ניתן לקבל רק מפיצ'ר זה לומד שצודק ב-60%.

c. *אורך הכותרת - KNB*
ראינו כי יש הבדל גדול באורך הכותרות בין "ישראל היום" לבין "הארץ" ולכן החלטנו לתת את אורך הכותרת כפיצ'ר. בדקנו אופציות שונות לכמות שכנים וראינו מה ממקסם את הלמידה ומצאנו כי 40 שכנים נותן את התוצאה הטובה ביותר.

d. *נקודה. - RFC*
שמנו לב כי באופן ניכר יש בכותרות של הארץ יותר נקודות (יש לציין כי בדקנו את כל הסימנים המיוחדים, ורק בנקודות ההבדל היה משמעותי). החלטנו לתת את הפיצ'ר הזה כפרט שיוסיף לחיזוי שלנו לגביי סיווג הכותרות.

e. *פוליטיקאים - RFC*
שמנו לב כי יש שמות של פוליטיקאים שחוזרים על עצמם יותר בעיתון מסויים מאחר. הוצאנו ממאגר נתונים את כל שמות הפוליטיקאים בארץ ובארה"ב ובנוסף הוספנו קובץ מיוחד שבעזרתו נתנו משקל נוסף לכותרות בהם טרמפ ונתניהו הופיעו (כיוון ששם ראינו הבדלים גדולים). ניתחנו את קבצי מסדי הנתונים הללו ויצרנו וקטורים שמייצגים כמה פעמים כל פוליטיקאי מופיע והשתמשנו בלמידה בשביל לקשר בין כמות האזכורים לסיווג. בנוסף נציין כי המשקל שנתנו לפוליטיקאים ה"מיוחדים" התבסס על מספר רב של ניסויים ולבסוף מתן המשקל האופטימלי שמקסם את אחוזי הלמידה.

5. <u>סיום הלמידה ומתן תחזית</u>
בשלב האחרון לקחנו את כל הפלטים שקיבלנו בשלב לעיל ושילבנו אותם על פי עיקרון ה-stacking. הלומד שלנו היה צריך ללמוד את המשקולות התקינות על מנת לקבל את הפרדיקציה הטובה ביותר תוך שילוב היתרונות בכל פיצ'ר ופיצ'ר. בצורה כזאת יכולנו להפיק את המיטב על ידי שילוב פיצ'רים ותכונות שונות. החלטנו לבסוף להשתמש בלומד Linear SVC שהפיק את התוצאות הטובות ביותר. בדקנו אופציות שונות לשכבה הסופית של הלמידה: RFC,KNF, SGD, שיטת הרוב קובע מהשכבה הראשונה ושיטת הרוב קובע בין כל הניסיונות של השכבה הסופית. כדי לקבל את התוצאה הטובה ביותר הרצנו את כל התהליך מספר רב של פעמים ותיעדנו מי ניפק את התיוגים המדויקים ביותר. ראינו כי Linear SVC קיבל את הציון הטוב ביותר בצורה משמעותית ולכן בחרנו ללמוד איתו את השכבה הסופית. בתוחלת עבור המודל הכולל המתואר לעיל קיבלנו בתוחלת תוצאה של 85.725%.

לסיום, תמונה של הצוות במהלך ההאקאתון: