מס' פרויקט	**שם הפרויקט**
2021 -01-179	Measuring Word Embedding Performance

מנחה אקדמי	**מנחה שותף**
פרופ' גלעד רביד	

חברי הצוות		
שקד אברהמי		
abrahamy@post.bgu.ac.il		

תקציר

חברת Exceed.AI הינה חברה המתמחה בפתרונות מבוססי בינה מלאכותית לממשק מול לקוחות. לחברה מספר מוצרים ובראשם ה AI Assistant שתפקידו לייעל תהליכי מכירה בכלל ואת התקשורת הראשונית מול לקוח פוטנציאלי בפרט. ה AI Assistant מציע את ההצעה הראשונית ללקוח פוטנציאלי, מנתח את תשובתו, מעריך את הסיכויים למכירה ומעביר לאנשי המכירות את הלקוחות לפי פוטנציאל המכירה הגבוה ביותר. הAI Assistant יכול לתקשר עם מספר רב של לקוחות בו זמנית לאורך כל שעות היום בעלות מינימלית ולכן מגדיל משמעותית את המכירות של משתמשיה. ניתוח טקסט המענה של לקוח וסיווגו לאשכול מסוים הינה משימת Text Classification השייכת לתחום ה natural language processing-NLP. Exceed.AI בחרה להשתמש במודלי Word Embedding כדי לנתח מיליים מלקוחות.

מטרת הפרויקט הינה לחקור ולפתח שיטה ייעילה ככל הניתן מבחינת משאבים לבחירת מודל ה Embedding עבור חברה כלשהי בכלל ו Exceed.AI בפרט.
הדרך המקובלת, הנפוצה כיום היא מומחה הקורא אלפי מיילים ומסווגם ידנית,הסיווג הוא מאוד איכותי אך לוקח זמן רב ויקר מאוד, בגלל אופי העבודה קשה למצוא מסווגים שיעשו זאת.

במהלך המחקר סקרתי את כלל השיטות המקובלות כיום להערכת מודל Word Embedding.
התמקדתי בשיטות לא מפוקחות (Unsupervised) ולאחר בחינת הנתונים החלטתי לפתח מדד מבוסס אישכול (Clustring).
המחקר התבסס על שני מקורות נתונים עיקריים, האחד מורכב מ- 63 אלף מיילים שנשלחו לחברות המשתמשות ב"AI Assistant ", עליו בוצע תהליך Data Processing מקיף. 1500 מיילים מתוכם היו מתוייגים ע"י צוות הData Science של Exceed. השני הוא "***google-news-300***" אשר שימש כסט ביקורת המורכב ממיליארדי משפטים מידיעות חדשותיות.
כשלב מקדים חילקתי בעזרת צוות הData Science של "Exceed" מילים שגורות בסט המיילים בעלות קשר סמנטי דומה ל4 קבוצות שונות. לכל מודל בעל היפר פרמטרים שונה ביצעתי מספר מבחני אישכול, לאחר מכן בדקתי עבור כל קבוצת מילים את התפלגותה על פני האשכולות השונים. הציפיה הייתה שאלגוריתם האישכול יתלכד עם הבנת חלוקת ה"מומחה" ויסווג קבוצת מילים באותו אשכול או במינימום אשכולות. תוצאות המדד תורגמו למדד אמפירי באמצעות אנטרופיה.
על מנת לבחון את הקשר בין מדדי האישכול למשימת ה" Text Classification" סווגו מיילים לקטגוריות השונות על סט האימון המתויג באמצעות רשת LSTM. עבור כל מודל בעל היפר פרמטרים שונה נמדדו Accuracy וה -Loss.
לאחר קבלת כלל התוצאות נבדק המתאם בין תוצאות האנטרופיה של המודלים למדדי ההצלחה השונים במשימת ה Text Classification.

תוצאות המחקר הראו כי עבור חלק מהמודלים קיים מתאם גבוהה בין המדדים.במחקר המשך יבדק האם באמצעות שינויים מינוריים בקוד לבצע מבחן לחלופות שונות של מודלים, להכריע מי מבינהם טוב יותר עבור דאטה סט ספציפי, כל זאת בחלקיק מהמשאבים שנדרשים בדרכים המסורתיות.

מילות מפתח:NLP, Clustering, linguistics, Machine learning, Word embedding