



השפעת אספקטים סמנטיים על הצלחה במימון המונים¹

אוראל בביוף
ביה"ס למנהל עסקים
אוניברסיטת בר אילן
orelbabayoff@gmail.com

און שחורי
ביה"ס למנהל עסקים
אוניברסיטת בר אילן
onn.shehory@biu.ac.il

תקציר

פלטפורמות מימון המונים מאפשרות ליזמים לפרסם פרויקטים ולגייס כספים למימושם. נשאלת השאלה, מהם הגורמים המשפיעים על הצלחת גיוס הכספים של פרויקטים. מחקרים קודמים בחנו השפעת גורמים שונים כגון מטרות הפרויקט ומשך הפרויקט על תוצאות קמפיניים לגיוס כספים. אנו מציגים מודל חדשני לחיזוי ההצלחה של פרויקטים למימון המונים בעמידה ביעדי המימון שלהם. המודל הבסיסי שלנו מתמקד בתכונות סמנטיות בלבד, וביצועיו דומים לאלה של מודלים קודמים. במודל מתקדם יותר שפיתחנו, אנו בוחנים הן מטא-נתונים של הפרויקטים והן את הסמנטיקה שלהם, ומציגים מחקר מקיף של הגורמים המשפיעים על הצלחת מימון ההמונים. בנוסף, מערך נתוני הפרויקטים למימון המונים שאנו מנתחים גדול באופן ניכר מן המאגרים עליהם דווח בספרות. לבסוף, אנו מראים שכאשר אנו משלבים סמנטיקה ומטא-נתונים, אנו מגיעים לדיוק בציון F1 של 96.2%. אנו משווים את הדיוק של המודל שלנו לדיוק של מודלים מחקרניים קודמים על ידי יישום השיטות שלהם על מערך הנתונים שלנו, ומדגימים דיוק גבוה יותר של המודל שלנו. בנוסף לתרומתנו המדעית, אנו מספקים המלצות מעשיות שעשויות להגדיל את סיכויי ההצלחה של מימון הפרויקטים.

מילות מפתח: מימון המונים, למידת מכונה, מודל חיזוי, תכונות סמנטיות.

הקדמה

בשנים האחרונות התפתחה שיטת מימון ההמונים כמנגנון פופולרי המאפשר לסוגים שונים של פרויקטים לקבל מימון. שלוש קבוצות שחקנים עיקריות מעורבות במימון המונים: יזמים (והפרויקטים שלהם), יחידים או קבוצות שתומכים בפרויקט, ופלטפורמות מקוונות למימון המונים [22]. האחרונות מפגישות יזמים ותומכים כדי לאפשר מימון הפרויקטים. קיימות פלטפורמות רבות למימון המונים, והפופולריות שבהן הן Kickstarter ו-Indiegogo [1]. עד שנת 2020 גויסו כ-34 מיליארד דולר בעזרת מימון המונים [7]. שיעור ההצלחה הממוצע של פרויקטים ממומנים למימון המונים Kickstarter הוא 37.5% [25]. במאמר זה אנו מתייחסים לפרויקט כאל הצלחה במימון אם הוא השיג

את יעד המימון שלו (כלומר, הוא גייס לפחות 100% מיעד המימון).

מחקרים רבים בחנו תכונות של פרויקטים, בנו מודלים של פרויקטים מוצלחים, וניסו לחזות הצלחה על סמך מודלים אלה. לדוגמה, [15] בחן את סוג המימון המבוקש ואת השפעתו על הצלחת המימון. ב- [28] נבחנה השפעת המעורבות החברתית של הפרויקט על הצלחה בגיוס המימון. ב- [5] נחקרו תכונות המטא-נתונים שמסבירות את הצלחת המימון, ועל סמך תכונות אלה פותח מודל לחיזוי הצלחת המימון. בעוד שמודלים בספרות מסתמכים על תכונות מטא-נתונים וניתוח נושאים כדי לחזות הצלחה במימון המונים, שימוש בתכונות סמנטיות לצורך חיזוי זה אינו שכיח. במחקר הנוכחי אנו נוקטים בגישה מקיפה לחיזוי הצלחה במימון, ולצורך כך אנו משלבים תכונות

¹ מבוסס על [2]

מבצעים ניתוח סמנטי מקיף של תיאור הפרויקט, כולל ניתוח נושאים באמצעות Latent Dirichlet Allocation (LDA), ניתוח שימוש במילים באמצעות Linguistic Inquiry and Word Count (LIWC) [17], וניתוחים נוספים, ששילובם מאפשר הפקת מאפיינים סמנטיים של הטקסט. שנית, אנו מציגים מודל חדשני המבוסס על תכונות סמנטיות בלבד, (להלן "המודל הסמנטי"), המנבא את הצלחתם של פרויקטים במימון המונים בעמידה ביעדי המימון שלהם. אנו מראים כי הדיוק של המודל הסמנטי (במונחים של ציון F ושל כמות הסיווגים הנכונים) דומה לדיוק של מודלי ייחוס המבוססים על תכונות מטא-נתונים כגון מספר העדכונים, מספר התמונות וכו'. שלישית, בהתבסס על התובנות שלנו לגבי סמנטיקה של פרויקטים, אנו מציגים מודל חיזוי נוסף המשלב סמנטיקה ומטא-נתונים של פרויקטים (להלן, המודל המשולב). הדיוק של המודל המשולב גבוה יותר מהדיוק שהתקבל במחקרים קודמים המשתמשים במילים אחרים [9, 21, 23, 27]. רביעית, אנו מראים שקיים מתאם גבוה ביותר בין תכונות סמנטיות הכוללות מילים אופנתיות (buzzwords), וכן תכונות LIWC, לבין הצלחתו של פרויקט בגיוס כספים.

בנוסף לתרומות אלה, מחקר זה מסתמך על מערך הנתונים הגדול ביותר ששימש עד כה לפיתוח מודלים המנבאים הצלחה במימון המונים. מערך הנתונים שבו נעשה שימוש כלל 160,125 רשומות פרויקטים משני אתרי מימון ההמונים הגדולים ביותר. גם לאחר סינון רשומות בעלות רלוונטיות נמוכה יותר, 111,273 רשומות פרויקט עדיין נכללו בניתוח. זה מאפשר לנו לספק מודל חיזוי מדויק מאוד עם רמות שגיאה נמוכות מאוד. בנוסף לתרומות המדעיות המפורסות לעיל, אנו מספקים סדרה של המלצות שעשויות להגדיל את סיכויי ההצלחה של מימון הפרויקט. המלצות אלה מבוססות על תוצאות המחקר שלנו בשילוב תוצאות מחקרים קודמים.

הניתוח הסמנטי שהוצג במחקר זה ישים מעבר לתחום הספציפי שנבחן בעבודה זו. באמצעות ניתוח כזה, ניתן לחקור את ההשפעה של תכונות סמנטיות של נתונים טקסטואליים על קהלי יעד שונים. במחקר שלנו נותחו הטקסטים המתארים את הפרויקט, וקהל היעד היה יחידים וקבוצות התומכים בפרויקט. תחום יישום אחר יכול להיות, למשל, ניתוח פוסטים ברשתות החברתיות

סמנטיות במודל החיזוי. אנו מחלצים מהטקסט של הפרויקט תכונות סמנטיות ומשתמשים בהן לבניית מודל ניבוי של הצלחת מימון. כדי להגדיל את דיוק התוצאות שלנו, אנו משלבים במודל שלנו תכונות ידועות בספרות המשפיעות על הצלחת המימון. גישה זו ממנפת את התוצאות הידועות ומשפרת אותן כדי לספק חיזוי הצלחה באיכות גבוהה. כפי שאנו מדגימים, שילוב תכונות סמנטיות בחיזוי הצלחה במימון אכן משפר את דיוק החיזוי באופן ניכר.

קיים מגוון רחב של פרויקטים המבקשים מימון המונים. הפרויקטים נבדלים זה מזה מבחינת סכום ההשקעה המבוקש וסוג התמורה המובטחת למשקיעים, ותכונות רבות נוספות. מקובל לסווג את הפרויקטים לארבעה סוגים עיקריים של מימון המונים [2, 15, 19]:

1. תגמול לא פיננסי: משקיעים תורמים בתמורה להטבות לא פיננסיות, כגון גישה מוקדמת ו/או חופשית למוצר או לשירות שפיתח הפרויקט.
2. הלוואה: המשקיעים מממנים פרויקטים כמלווים ומקבלים ריבית על השקעתם.
3. שותפות הונית: המשקיעים מקבלים ניירות ערך של החברה המוקמת בפרויקט.
4. תרומה (מימון ללא תגמול): מיועד למטרות צדקה או לגיוס כספים לפרויקטים חברתיים.

המטרה העיקרית של מחקר זה היא לאתר מאפיינים, ובפרט במאפיינים סמנטיים, המשפיעים על הצלחתו של פרויקט לעמוד ביעד המימון שלו באמצעות מימון המונים. מטרה נוספת היא לענות על השאלות הבאות: בהינתן מערך נתונים של פרטי פרויקטים במימון המונים הכולל טקסטים (כגון תיאור פרויקט), נתונים לא טקסטואליים (כגון קטגוריית הפרויקט, סכום המימון הנדרש) ומטא-נתונים, האם אנו יכולים לחזות את הצלחתם של פרויקטים בעמידה ביעדי המימון שלהם? באיזו רמה של דיוק? כיצד אלגוריתמי למידה שונים ישפיעו על דיוק זה? השגת יעדי מחקר אלה תספק תובנות חדשות על הקשר שבין נתוני פרויקט מימון המונים לבין הצלחת המימון שלו. בנוסף, תובנות אלה יכולות לפתוח בפני יזמים דרכים נוספות לשיפור הסיכויים לעמידה ביעדי המימון שלהם.

המחקר מציג ארבע תרומות מדעיות עיקריות. ראשית, אנו מדגימים את חשיבותן של התכונות הסמנטיות של פרויקט בעמידה ביעדי גיוס הכספים שלו. לשם כך, אנו

שימוש ב-LIWC

כדי לחלץ תכונות מהטקסט, השתמשנו ב-LIWC, הכולל תוכנה לניתוח טקסט וקבוצה של מילונים נושאיים מובנים [17]. ניתוח LIWC מודד את שכיחותן של מילים ממילון מסוים בטקסט הנבדק. המדד הוא ערך מספרי בתחום [1..0] המשקף את המידה שבה הטקסט המנותח קשור לנושאי המילון. בפרט, הניתוח עובד כדלקמן. N_T הוא מספר המילים בטקסט T , ו- $M_D = \{w_1, \dots, w_n\}$ הוא קבוצת המילים במילון D . בהתאם, N_w הוא מספר המופעים של המילה w מ- M_D ב- T . $S_D = \sum_{i=1..n} N_{wi} \cdot T$ הוא הסכום של N_w על כל המילים w ב- M_D . הפלט של LIWC הוא $V_D = S_D / N_T$. זהו ערך התכונה המתאימה למילון D . למעשה, V_D מודד את תדירות הופעת המילים (ושורשיהן) ממילון מסוים D בטקסט הנבדק.

מחקרים קודמים על מימון המונים השתמשו ב-LIWC כדי לנתח את התיאור הטקסטואלי של פרויקטים והגיעו לדיוקים גבוהים במודל (למשל, ציון $F > 0.8$) [27]. הבדל בולט בין המחקר הנוכחי לבין מחקרים אלו הוא שהמספר והמגוון של תכונות שנותחו במחקר שלנו גדולים משמעותית לעומת עבודות קודמות, וגם מערך הנתונים שלנו גדול יותר באופן ניכר. כתוצאה מכך, המודל שלנו מגיע לדיוק גבוה יותר.

הקצאת דיריכלה סמויה (LDA)

שיטות מידול נושאים משמשות במשימות כריית טקסט כדי לחלץ נושאים מטקסט. הקצאת דיריכלה סמויה (LDA) היא שיטת מידול נושאים נפוצה [27]. אלגוריתם LDA מתייחס לכל מסמך כאוסף של נושאים, כאשר כל מילה במסמך שייכת לנושא אחד או למספר נושאים. לכל מילה יש משקל המבטא את חשיבותה בהקשר של הנושא אליו היא משויכת. נושא בא לידי ביטוי על ידי קבוצת המילים השייכות לו ומשקליהן. ערכו של נושא יהיה סכום משקלים האלה. לאחר שהאלגוריתם זיהה מספר נושאים כנדרש, הוא מסדר מחדש את התפלגות הנושאים בתוך המסמכים ואת התפלגות מילות המפתח בתוך הנושאים כדי לקבל תצורה טובה של התפלגות הנושאים למילות מפתח [27].

מחקרים קודמים על מימון המונים השתמשו ב-LDA כדי לבצע ניתוח נושאים על הטקסט של עדכונים פרויקטים, אם כי ניתוח זה בוצע על מערך נתונים קטן

כדי לבחון את רמת התמיכה שמפגינים עוקבים בארגון שאחריו הם עוקבים.

המאמר ממשיך כדלקמן. פרק המבוא מציג עבודות קודמות, את המוטיבציה למחקר הנוכחי, ואת מטרותיו. פרק הרקע מציג רקע במימון המונים ובניתוח סמנטי. פרק המתודולוגיה מתמקד במתודולוגיית המחקר, כולל איסוף הנתונים, עיבוד מקדים, וחילוף תכונות. פרק התוצאות מציג את הבחינה האמפירית של מודל המחקר ודן בתוצאות. הוא גם משווה את המודלים שלנו למודלים במחקרים קודמים. לבסוף, בפרק המסקנות אנו מסכמים, דנים בממצאים העיקריים ומצביעים על כיוונים עתידיים.

רקע

בפרק זה אנו דנים בתכונות פרויקטים עליהן נבנים מודלי חיזוי ובמדדי הדיוק של המודלים, ומדגישים את השוני בין מחקרים קודמים למחקר הנוכחי. לצורך ניתוח סמנטי, נדרשים הנתונים הטקסטואליים של הפרויקטים. לפיכך, מקור נתונים מרכזי המשמש במחקר זה הוא פוסטים של פרויקטים במימון המונים. פוסטים כאלה כוללים בדרך כלל מספר מרכיבים, ביניהם: כותרת, כותרת משנה, קטגוריה, שאלות נפוצות, תיאור הפרויקט, וידאו, מטרות המימון, זמן להשגת המטרה, הסכום שגויס, פרטי קשר, עדכונים, תומכים, דרכי תמיכה ושמות היזמים או החברה. הניתוח הסמנטי שלנו כולל ניתוח נושאים באמצעות LDA, ניתוח שימוש במילים באמצעות LIWC ותכונות חדשניות כולל שימוש במילים אופנתיות, מילים המבטאות רגשות, ומילות הסבר.

מילים אופנתיות (Buzzwords)

בניתוח הפוסטים שערכנו נבחן השימוש במילים אופנתיות [4]. למרות שמחקרים קודמים בחנו שימוש במילים אופנתיות בטקסטים של פרויקטים, הקשר בין הצלחת המימון לבין מילים אופנתיות ששימשו לתיאור הפרויקט לא נבחן עד כה. זוהי תרומה ייחודית של המחקר שלנו. קבוצת המילים אופנתיות המשמשת במחקר זה מכילה מילים מקטגוריות שונות כגון חינוך, עסקים, מכירות ושיווק, מדע וטכנולוגיה, ופוליטיקה ואקטואליה. לדוגמה, מילים אופנתיות מתחום הטכנולוגיה עשויות להיות ביג דאטה, מחשוב ענן, קוד פתוח וכו'.

הפרויקטים של היזם שהושלמו בהצלחה [5, 6, 10, 14, 20, 21, 24, 26].

תכונות סמנטיות נוספות

בנוסף לנושאים שזוהו על ידי LIWC ו-LDA, סקרנו מחקרים בתחום הניתוח הטקסטואלי כדי לאתר נושאים נוספים שלהם פוטנציאל למתאם גבוה להצלחה במימון. הנושאים הבאים אותרו:

• מילות הסבר: example, explain, for instance, i.e., mean, in other words, in that, that is [8].

• מילות רגש: angry, annoyed, afraid, awkward, affectionate, anxious, alarmed, awed, aggravated, amazed וכו'. רשימת מילים זו התקבלה מאיחוד מילוני LIWC הכוללים מילות רגש [12].

מדדי ביצועים

השתמשנו במדדים הבאים כדי למדוד את הביצועים של המודלים שלנו. נכונות (Precision) הוא היחס בין מספר הרשומות שערך חיובי וגם נחזו כחיוביות לבין סך הרשומות שנחזו כחיוביות. איחזור (Recall) הוא היחס בין מספר הרשומות שערך חיובי וגם נחזו כחיוביות לבין כל הרשומות באותה מחלקה. F-score (מסומן גם F_1) הוא הממוצע ההרמוני המשוקלל של Precision וה Recall של המודל. מדדים אלה מחושבים באופן הבא:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

כאשר t_p הוא מספר התחזיות החיוביות האמיתיות. f_n ו- f_p הם מספר התחזיות השליליות הכוזבות והחיוביות הכוזבות, בהתאמה.

בהרבה וספציפי מאוד בהשוואה למחקר שלנו [27]. מחקר אחר השתמש ב-LDA כדי לנתח אתרי מימון המונים סיניים והגיע לרמת דיוק (ציון F) של 86.7%. במחקר שלנו, השתמשנו ב-LDA כדי לחלץ תכונות מתיאורי פרויקטים, כאשר כל נושא שימש כתכונה במודל. השתמשנו ביישום LDA של חבילת Gensim Python (מתוך github.com), הידוע כמהיר יותר מיישומים אחרים, וכמייצר הפרדת נושאים טובה יותר [3].

לצורך ניתוח LDA ביצענו עיבוד מקדים. ראשית, ארגנו כל משפט לרשימת מילים, תוך הסרת סימני פיסוק, תווים מיותרים, כתובות דוא"ל ושמות. לאחר מכן הסרנו מילות העצירה, איתרנו bigrams (שתי מילים המופיעות לעתים קרובות יחד במסמך) ובצענו lemmatization.

שלושת הקלטים העיקריים למודל הנושאים של LDA הם המילון, הקורפוס ומספר הנושאים. מבוכה (perplexity) של המודל וקוהרנטיות של נושאים הינם מדדים המאפשרים לשפוט את איכותו של מודל הנושאים [3]. כדי למנוע התאמת יתר, שינינו את מספר הנושאים מ-10 ל-50 ובחרנו את המספר הקטן ביותר שהוביל לקוהרנטיות הגבוהה ביותר.

הרצנו את ה-LDA בנפרד על שלושה מערכי נתונים והפקנו שלוש קבוצות נושאים – אחת לכל מערך. הקלט ל-LDA היה תיאורים טקסטואליים של הפרויקטים במערך נתונים. הפלט היה קבוצה של נושאים ואחוז התרומה של כל נושא לכל טקסט קלט. במחקר שלנו, כל נושא שימש כתכונה לפיתוח מודל חיזוי, וערך התכונה היה אחוז התרומה של נושא. נתקבלו 30 נושאים שהניבו קוהרנטיות גבוהה עבור מערך הנתונים All_D, 15 עבור מערך הנתונים Tech_D ו-15 עבור מערך הנתונים Market_D.

מטא-נתונים

בנוסף, שילבנו תכונות מטא-נתונים הידועות כמשפיעות על הצלחת המימון אותן חילצנו מפוסטים של פרויקטים. התכונות בהן השתמשנו לניתוח שלנו כללו את מספר התמונות, מספר הסרטונים, מספר העדכונים, יעד המימון והזמן להשגת היעד, מספר הפרויקטים שנוצרו בעבר על ידי היזם ומספר

מתודולוגיה

על מנת לעמוד ביעדי מחקר זה, פיתחנו שיטה לגילוי ידע הכוללת ארבעה מרכיבים עיקריים: השגת מערכי נתונים והפעלת עיבוד מקדים; בחירת תכונות; ניתוח נתונים מבוסס מודלי למידת מכונה; ולבסוף, בחינת הידע שחולץ על ידי המודלים.

השתמשנו במערך נתונים של 50,000 פרויקטים מ-Kickstarter ו-50,000 פרויקטים מ-Indiegogo. בדומה למחקרים אחרים בתחום, סיננו פרויקטים שלהם פחות מ-10 שורות תיאור. מטרת הסינון היא להימנע מפגיעה בדיוק של הניתוח הסמנטי [27]. כמו כן, השמטנו ממערך הנתונים פרויקטים שתאריך היעד שלהם היה עתידי. לאחר הסינון נותרו 73,580 פרויקטים: 42,530 שהשיגו את יעד המימון שלהם (הצלחה = 1), ו-31,050 פרויקטים שלא הצליחו להשיג את יעד המימון שלהם (הצלחה = 0).

לאחר סינון מערך הנתונים, השתמשנו ב-Beautifulsup (Python package) כדי להשיג תכונות מטא-נתונים נוספות (כגון מספר עדכונים שבוצעו). עבור כל רשומה במערך הנתונים, קיבלנו ערכים של חמש תכונות מטא-נתונים וכ-120 תכונות סמנטיות, כולל מילים אופנתיות, פלטים של LIWC, מילות רגש, מילות הסבר ופלטים של LDA. ערכי התכונות חושבו כמתואר לעיל. למניעת כפילויות, נושאים שהופקו על ידי LDA שלהם חפיפה גבוהה עם נושאים שהופקו על ידי LIWC הוסרו. ערכנו עיבוד מקדמי של מערך הנתונים בשיטות סטנדרטיות (לדוגמה, המרה של משתנה אורדינלי למשתנה מספרי, או נורמליזציה של מערך נתונים) כדי להתאים את הנתונים לדרישות הקלט של אלגוריתמי ניתוח הנתונים שבהם השתמשנו לפיתוח המודל.

כדי לחקור את הקשר בין תכונות לבין קטגוריית הפרויקט, בנינו שלושה מערכי נתונים:

- All_D: הכולל את כל הפרויקטים ללא קשר לקטגוריה שלהם.
- Tech_D: כולל פרויקטים טכנולוגיים בלבד.
- Market_D: כולל פרויקטים בנושאי שיווק ועסקים בלבד.

חלוקה זו נועדה לתת מענה למטרות הניתוח הבאות. ראשית, היא מאפשרת שיפור ברמת הקוהרנטיות של

נושאי ה-LDA. שנית, היא מאפשרת לנו לקבוע אם קטגוריה ספציפית משפיעה על התכונות שנמצאות במתאם גבוה עם הצלחת המימון. שלישית, כדי לחקור את המתאם בין מילים אופנתיות להצלחה במימון, עדיף להפריד מערכי נתונים לתת-תחומים ספציפיים (רוב המילים האופנתיות שהשתמשו בהן אכן נמצאות בתת-תחומים אלה).

כדי לבחון את הדמיון בין מערכי הנתונים השתמשנו ב-MANOVA (ניתוח רב-משתני של שונות). התוצאות מראות כי ההפרדה לפי קטגוריה גורמת למובהקות סטטיסטית של ממוצע ההבדלים בין התכונות (p -value = 0.05). חישבנו מתאם פירסון בין כל התכונות שחקרנו לבין הצלחה במימון. כל המודלים וערכי המתאם חושבו, בנפרד, ביחס לשלושת מערכי הנתונים.

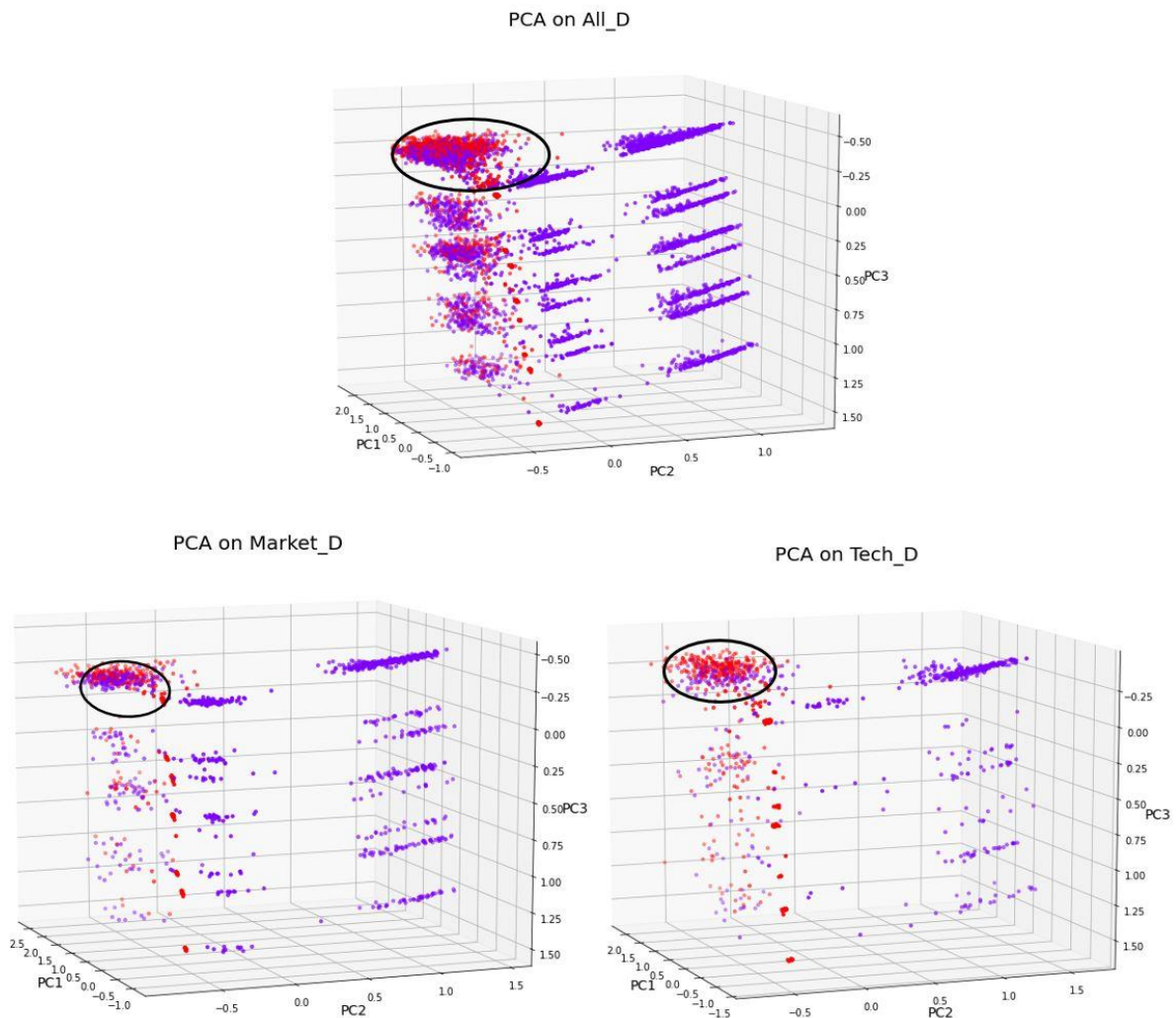
לצורך אילוסטרציה של מרחב התכונות, היה עלינו לצמצם את המרחב הרב-ממדי לתלת-ממדי. לפיכך, השתמשנו בניתוח רכיבים ראשיים (Principal Component Analysis) כדי לזהות את התכונות המשפיעות ביותר [11, 29]. כך, עבור כל מערך נתונים, מרחב התכונות המקורי צומצם לשלושה רכיבים עיקריים, PC1, PC2, PC3, כפי שמודגם באיור 1. האיור מציג את התפלגות שלוש התכונות המשפיעות ביותר במערכי הנתונים All_D, Market_D Tech_D. כל נקודה באיור מייצגת רשומת פרויקט במערך הנתונים. נקודות בכחול מייצגות פרויקטים ללא הצלחה במימון ונקודות באדום מייצגות פרויקטים שהצליחו במימון. רוב הרשומות של הפרויקטים שמומנו בהצלחה ממוקמות בתחום שבו $1 < PC1 < 2$, $-1 < PC2 < 0$, $PC3 < 0.5$. באיור, תחום זה מוקף באליפסה שחורה.

כדי לזהות את קבוצת התכונות המשמעותית ביותר (Most Significant Set of Features - MSSF) שלה ההשפעה הגבוהה ביותר על הצלחה במימון, אנו משתמשים באלגוריתם CFS (Correlation-based Feature Selection). האלגוריתם בוחר תת-קבוצה של תכונות על בסיס יכולת הניבוי האינדיבידואלית של כל תכונה יחד עם מידת היתירות ביניהן. הפעלנו את האלגוריתם על כל מערך נתונים בנפרד.

כאמור, פיתחנו שני מודלי סיווג בינאריים כדי לחזות את הצלחת המימון. תוויות המחלקות הן 1 (הצלחה

במימון) ו-0 (כישלון במימון). כדי לפתח את המודלים השתמשנו באלגוריתמי סיווג נפוצים (יער אקראי, עצי החלטה, DNN) וכן באלגוריתמים מובילים כמו LightGBM [13]. פיתוח המודלים נערך באמצעות חבילות ה-Python הבאות: scikit-learn, scikit-feature ו-LightGBM.

איור 1. תוצאות PCA על שלשת מערכי הנתונים



כך פיתחנו את המודל המשולב, שהחידוש שלו טמון בשילוב של תכונות סמנטיות ומטא-נתונים. עבור כל אחד משני המודלים, חישבנו את ציון ה-F כמדד לדיוק המודל.

לבדיקת ביצועי המודל השתמשנו במבחן 10-fold cross-validation. מבחן זה שימש לבדיקת הביצועים

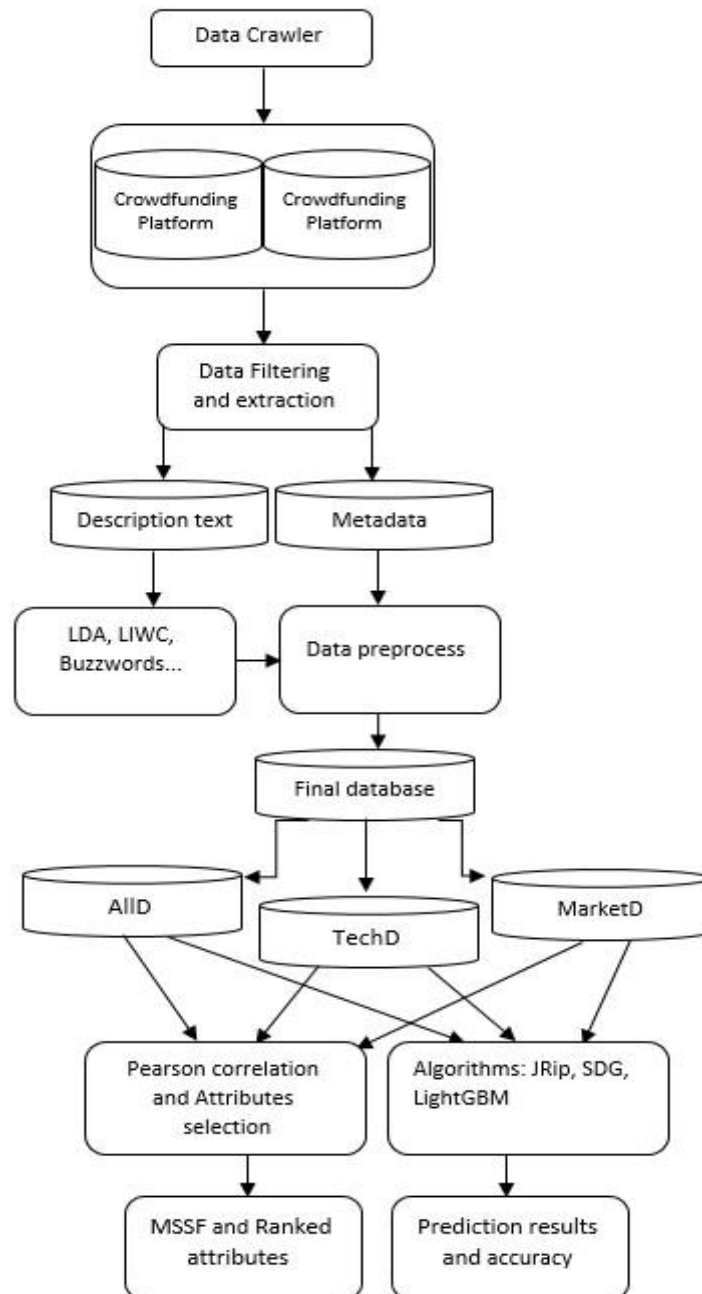
נזכיר שאחת המטרות שלנו הייתה לפתח מודל הסמנטי שהדיוק שלו דומה לדיוק של מודלים קיימים. לשם כך, פיתחנו מודל שתכונותיו הן רק תכונות סמנטיות. הקלטים של מודל זה מורכבים ממילים אופנתיות, מילות הסבר, מילות רגש ומהנושאים המופקים מ-LIWC ו-LDA. לאחר שהמודל הסמנטי פותח ונבדק, שמנו לנו למטרה לשפר עוד יותר את דיוק המודל. לשם

לאשש את תוצאות המחקר שלנו, השווינו את דיוק של המודל שלנו לדיוק של מודלים מחקרניים קודמים על ידי יישום השיטות שלהם על מערך הנתונים שלנו.

לסיכום פרק זה, איור 2 מציג תרשים זרימה של המתודולוגיה.

של מודלים דומים לשלנו [18, 24, 28]. לכל fold, מערך הנתונים חולק באופן אקראי ל-10 חלקים, ואז בוצעו 10 ניסויים, כאשר תשעה חלקים שימשו כנתוני אימון למודל כדי לחזות את החלק הנותר. לכל fold, אלגוריתם ה-LDA אומן על אותם נתוני אימון. כדי

איור 2. תרשים זרימה של המתודולוגיה



תוצאות

חילוץ ובחירה של תכונות

בפרק זה אנו מספקים פרטים על הגדרת הנתונים, השימוש באלגוריתם LDA, ניתוח המתאם בין התכונות, תהליך בחירת התכונות ופיתוח מודל החיזוי.

התכונות המשפיעות ביותר על הצלחת מימון

לצורך מציאת התכונות המשפיעות ביותר על הצלחת מימון השתמשנו בשתי שיטות. הראשונה היא מתאם

פירסון בין התכונות הבלתי תלויות לבין התכונה התלויה (כלומר, הצלחת המימון). השנייה היא אלגוריתם CFS.

תכונות שלהן מתאם גבוה עם הצלחה במימון

בטבלה 1, עבור כל אחד משלושת מערכי הנתונים, אנו מפרטים את 10 התכונות שעבורן נמצאו ערכי המתאם הגבוהים ביותר להצלחה במימון. הסמנטיקה של תכונות המסומנות ב-* מצויה בתיעוד של LIWC.

טבלה 1. מתאם של תכונות עם הצלחה במימון

	מתאם	Market_D	מתאם	Tech_D	מתאם	All_D
1	0.717	punc*	0.718	achieve*	0.758	achieve*
2	0.696	achieve*	0.677	percept*	0.668	punc*
3	0.659	preps*	0.675	feelings_count	0.625	preps*
4	0.54	work*	0.663	WC*	0.599	feelings_count
5	0.495	see*	0.644	see*	0.595	percept*
6	0.494	present *	0.61	discrep*	0.592	work*
7	0.492	feelings_count	0.574	updates	0.583	hear*
8	0.478	Instagram_link	0.556	work*	0.582	see*
9	0.475	buzz_count	0.53	punc*	0.56	discrep*
10	0.468	number*	0.52	buzz_count	0.519	buzz_count

בחירת תכונות

השתמשנו באלגוריתם CFS כדי למצוא MSSF בכל מערך נתונים. אלגוריתם CFS מעריך תת-קבוצות של תכונות בהתבסס על יכולת הניבוי האינדיבידואלית של

כל תכונה יחד עם מידת היתירות ביניהן. תוצאות האלגוריתם מוצגות בטבלה 2.

טבלה 2. קבוצות התכונות המשמעותיות ביותר (MSSF) במערכי הנתונים

All_D	Tech_D	Market_D
achieve*	backed	funding goal
updates	funding goal	updates
punc* (all punctuation)	buzz_count	buzz_count
emoticon*	updates	work*
funding goal	FAQ (# FAQ in project's page)	created (# created projects)
backed (# backed projects)	project duration	punc*
project duration		

המחקר שלנו היא שתכונות הקשורות להצלחה במימון תלויות בקטגוריית הפרויקט.

מודלים לחיזוי הצלחה במימון

המודל הסמנטי

לפיתוח המודל הסמנטי השתמשנו בתכונות סמנטיות כקלט. השתמשנו במספר אלגוריתמים של למידת מכונה ובהם SVM, J48, Random Forest, LightGBM, SDG, DNN ועוד. טבלה 3 מציגה את מדדי הדיוק של המודל הסמנטי שהופעל על מערך הנתונים All_D.

מהתוצאות לעיל אנו מסיקים כי קיים מגוון של תכונות סמנטיות שלהן מתאם גבוה להצלחה במימון. ערכי מתאם דומים נמצאו גם במחקרים קודמים שלא התמקדו בתכונות סמנטיות [14, 21]. התכונות הסמנטיות buzzwords ו-LIWC הן בין התכונות שלהן מתאם גבוה בכל מערכי הנתונים. בתוצאות ניתוח המתאם ניתן להבחין כי בין 10 התכונות המובילות, כ- 70% מהתכונות זהות בין מערכי נתונים. לדוגמה, התכונות feelings_count ו-buzz_count הן בין 10 התכונות המובילות בשלושת המערכים. לעומת זאת, הפרמטר WC הוא ייחודי ברשימת 10 המובילים ב-Tech_D ומספר הפרמטרים ייחודי ברשימת 10 המובילים ב-Market_D. מסקנה חשובה וייחודית של

טבלה 3. ביצועי המודל הסמנטי

Algorithm	F-score	Precision	Recall	Accuracy
LightGBM	91.1%	92.3%	89.9%	90 %
SDG	89.3%	90.5%	88.1%	88.5%

המודל המשולב

בשלושת מערכי הנתונים, תוך שימוש בכל התכונות.
טבלאות 4-6 מציגות את הביצועים של המודלים
הטובים ביותר (לפי ציון F).

כדי לפתח את המודל המשולב, השתמשנו באותם
אלגוריתמי סיווג ששימשו לפיתוח המודל הסמנטי

טבלה 4. ביצועי המודל המשולב – All_D

Algorithm	F-score	Precision	Recall	Accuracy
LightGBM	96.2%	97%	95.4%	96.2%
SDG	95.3%	95.5%	95.1%	95.1%

טבלה 5. ביצועי המודל המשולב – Tech_D

Algorithm	F-score	Precision	Recall	Accuracy
LightGBM	94.8%	94.8%	94.8%	94.2%
Random	93.9%	91.7%	96.2%	92.2%

טבלה 6. ביצועי המודל המשולב – Market_D

Algorithm	F-score	Precision	Recall	Accuracy
LightGBM	95.6%	95.9%	95.3%	95.3%
SDG	95.2%	95.4%	95%	95.4%

grid search קבוצה של היפר-פרמטרים, ואלה
שימשו לאימון המודל שלנו. בפרט, נעשה שימוש בהיפר-
פרמטרים הבאים עבור מערך הנתונים All_D:

• SDG: גודל אצווה = 100, epochs = 500, קצב למידה
= 0.01, פונקציית loss = hinge loss (SVM), למבדה
(קבוע רגולריזציה) = 0.0001.

השתמשנו באלגוריתם grid search לצורך
אופטימיזציה של ההיפר-פרמטרים שאיתם בוצע כל
אלגוריתם (לדוגמה, עבור SDG ביצענו אופטימיזציה
של גודל האצווה, מספר ה-epochs, וכו'). בנוסף
לאלגוריתם grid search, השתמשנו באלגוריתם
random search וקיבלנו דיוקים דומים. אלגוריתם

ביותר. אנו מציינים את המודל שהתאמן על תכונות אלה בשם מודל ה-LDA. מחקר נוסף [10] השתמש רק בתכונות מטא-נתונים כדי לפתח את מודל החיזוי. אנו מציינים את המודל שהתאמן על תכונות אלה בשם מודל המטא-נתונים.

המחקר שלנו בדק אם קבוצת התכונות ששימשו לחיזוי ומערך הנתונים שעליו יושמה הלמידה מספקים מודל מדויק יותר ממודלים קודמים (במונחי ציון F). לשם כך בחנו את ההשפעה של אלגוריתמי הלמידה, של קבוצות התכונות ושל מערך הנתונים על ציון ה-F המתקבל. בהתאם לכך, אימנו את מודל ה-LDA ואת מודל המטא-נתונים עם האלגוריתמים ששימשו במחקרים שצוינו לעיל, ועם אלגוריתמים נוספים, כולל SVM, 48J, Random Forest, LightGBM, SDG, ו-DNN.

אימון זה בוצע על All_D עם 10-fold cross-validation. ניסוי מקיף זה הוביל למסקנה כי לא אלגוריתמי הלמידה ולא מערך הנתונים הם מקור להבדלים משמעותיים בציון F. המשפיע העיקרי על הבדלים כאלה הוא קבוצת התכונות. איור 3 מציג את הערכי ציון F הגבוהים ביותר של מודל ה-LDA, מודל המטא-נתונים, המודל הסמנטי והמודל המשולב. האזור מראה שהמודל שפיתחנו הוא בעל הדיוק הגבוה ביותר. הדיוק של המודל הסמנטי דומה לדיוק של מודל ה-LDA וגבוה מהדיוק של מודל המטא-נתונים.

• Random Forest : מספר העצים = 20, עומק מקסימלי = אינסוף, מספר מרבי של תכונות = לא מוגבל.

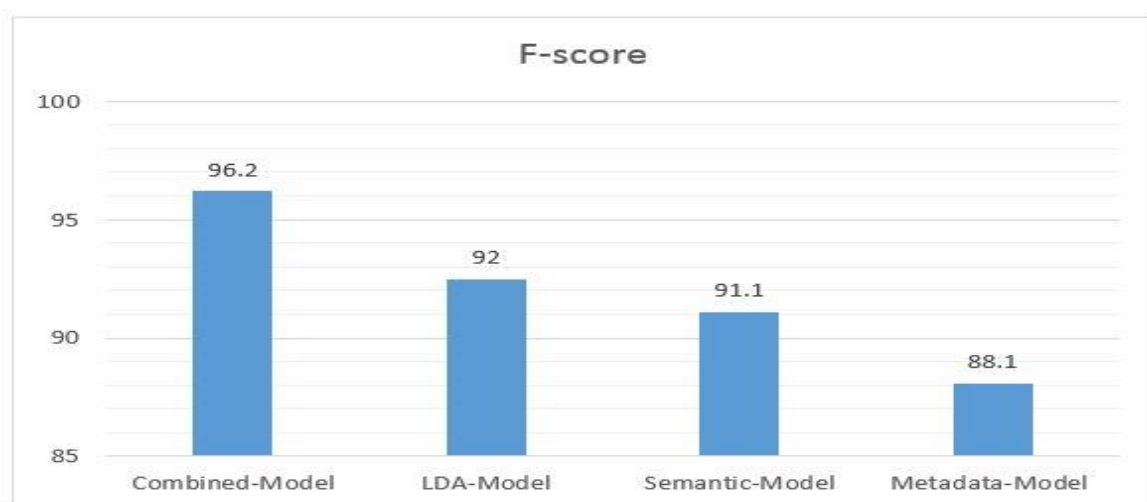
• LightGBM : מספר עלים = 255, מספר איטרציות מגבירות = 5000, מספר מינימלי של נתונים בעלה אחד = 10, קצב למידה = 0.01, מספר מרבי של תאים = 300.

כפי שניתן לראות לעיל, המודל הסמנטי שפיתחנו דומה בדיוקו למודלים ידועים המסתמכים על מטא-נתונים, ומציג רמת דיוק של 91.2%. יתר על כן, כפי שמוצג בטבלה 4, שילוב המודל הסמנטי עם מודל המטא נתונים הוביל לרמת דיוק של 96.2%. בנוסף, כפי שניתן לראות בטבלאות 4-6, אלגוריתם LightGBM הציג רמת דיוק גבוהה עבור כל שלושת מערכי הנתונים. בפרט, עם LightGBM, הדיוק של מודל החיזוי הוא יותר מ-94% עבור כל מערכי הנתונים.

השוואת המודל שלנו למודלים קודמים

כדי להבין טוב יותר את משמעות התוצאות שלנו, השווינו אותן לתוצאות של מחקרים קודמים. לשם כך, יישמנו שני מודלים ממחקרים קודמים ואימנו אותם על מערך הנתונים All_D כדי להשוות את דיוקם לאלו של המודל הסמנטי והמודל המשולב.

מחקר ב- [27] מציג המודל שאומן על 80 תכונות LDA שחולצו מתיאורי פרויקטים. בהסתמך על השיטה שלהם, ניסינו מספר שונה של תכונות בטווח של 5 עד 100 ומצאנו ש-30 תכונות ביאות לביצועים הגבוהים



איור 3. השוואת דיוק המודלים

מסקנות

שמחקרים קודמים בחנו. בנוסף, הראינו שתכונות שלהן מתאם גבוה עם הצלחה במימון תלויות בקטגוריית הפרויקט.

מנקודת מבט מעשית, תוצאות המחקר שלנו רלוונטיות מאוד לגיוסי כספים באמצעות פלטפורמות אינטרנטיות למימון המונים. בנוסף לתרומות המדעיות המפורטות לעיל, ניתן להציע סדרה של המלצות שעשויות להגדיל את סיכויי ההצלחה של מימון הפרויקט. המלצות אלה מבוססות על התכונות שמנינו לעיל, ועל תכונות משפיעות ממחקרים קודמים. הראינו כי הקטגוריה של הפרויקט משפיעה על התכונות שיש להן מתאם גבוה עם הצלחת המימון. אנו מציעים מספר המלצות לפרויקטים טכנולוגיים, כדלקמן:

- עדכון פרטי הפרויקט במהלך תקופת המימון.
- הכללת מילות רגש בפוסט של הפרויקט.
- הכללת מילים אופנתיות בתיאור הפרויקט.

בעוד שמאמר זה הגיע לדיוקי חיזוי גבוהים מאוד של הצלחת מימון, מחקר עתידי יכול לשפר עוד יותר את הדיוק על ידי התחשבות במאפייני תמונות ווידאו והסמנטיקה שלהם. שיפורים נוספים בביצועי המודל יכולים להיות מושגים באמצעות אלגוריתמים חדשניים לבחירת תכונות, למשל [30].

בשנים האחרונות, פלטפורמות מימון המונים כמו Kickstarter ו-Indiegogo מאפשרות ליזמים להציג את הפרויקטים ולגייס את הכספים הנדרשים להם. השאלה כיצד מאפיינים שונים של הצגת הפרויקט יכולים להגדיל את הסיכוי למימון מוצלח של הפרויקט היא חשובה. מחקרים קודמים זיהו תכונות מטא-נתונים וניתוח נושאים כבסיס לחיזוי הצלחה במימון המונים, מעט מאוד מחקרים בחנו תכונות סמנטיות בהקשר זה.

במחקר זה התייחסנו לחוסר הזה. התמקדנו בניתוח היבטים סמנטיים של פוסטים בפרויקטים כדי לשפר את הדיוק של תחזיות הצלחה במימון. בנינו שיטה לניתוח טקסטים של פרויקטים ופיתחנו מודל לניתוח וחיזוי הצלחה במימון המונים. תחילה, פיתחנו מודל חדשני המבוסס על תכונות סמנטיות בלבד והגענו לרמת דיוק דומה לזו של מחקרים קודמים. בנוסף, פיתחנו מודל חיזוי עם ציון F מרשים של 96.2%, תוך התמקדות הן בהיבטים ספציפיים לפרויקט והן בסמנטיקה של תיאורי הפרויקט. למיטב ידיעתנו, מחקר זה הוא הראשון שחוקר את הקשר בין הצלחה במימון לבין מילים אופנתיות. אנו מראים שתכונות המילים האופנתיות היא בין התכונות שנמצאות במתאם גבוה להצלחה במימון בהשוואה הן לפרמטרים שאנו בחנו והן לפרמטרים

ביבליוגרפיה

1. Digital Trends, 2019. <https://www.digitaltrends.com/cool-tech/best-crowdfunding-sites/> (accessed 1.12.2019).
2. Babayoff, O., Shehory, O., 2022. The role of semantics in the success of crowdfunding projects. PLoS ONE 17(2).
3. Beaulieu, T., Sarker, S., Sarker, S., 2015. A conceptual framework for understanding crowdfunding. Communications of the Association for Information Systems 37, 1-31.
4. Blei, D.M., Ng, A. Y., & Jordan, M.I., 2003. Latent Dirichlet allocation. Mach. Learn. Res. 3(Jan) 993-1022.

5. Buzzword, 2015. <https://www.merriam-webster.com/dictionary/buzzword>, (accessed 3 February 2015).
6. Cordova, A., Dolci, J., & Gianfrate, G., 2015. The determinants of crowdfunding success: Evidence from technology projects. *Procedia - Soc. Behav. Sci.* 181, 115-1254.
7. Frydrych, D., Bock, A.J., Kinder, T., Koeck, B., 2014. Exploring entrepreneurial legitimacy in reward-based crowdfunding. *Venture Capital* 16, 247-269.
8. Fundly. 2020. <https://blog.fundly.com/crowdfunding-statistics/> (accessed day month year).
9. Ghose, A., & Ipeirotis, G.P. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and review characteristics. *IEEE Trans. Knowl. Data Eng.* 23(10), 1498-1512.
10. Greenberg, M. D., Pardo, B., Hariharan, K., & Gerber, E., 2013. Crowdfunding support tools: Predicting success & failure. *CHI '13 Ext. Abstr. Hum. Factors in Comput. Syst. ACM, New York*, pp. 1815-1820.
11. Jascha-Alexander, K., & Michael, S., 2015. Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms. *Twenty-Third European Conference on Inform. Syst. (ECIS) paper 106*.
12. Jolliffe, I.T., 2002. *Principal Component Analysis and Factor Analysis* Principal Component Analysis. Springer, New York.
13. Kahn, J.H., Tobin, R.M., Massey, A.E., & Anderson, J. A., 2007. Measuring emotional expression with the linguistic inquiry and word count. *Amer. J. Psychol.* 120(2), 263-286.
14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pp. 3146–3154.
15. Kim, T., Por, M. H., & Yang, S.-B., 2017. Winning the crowd in online fundraising platforms: The roles of founder and project features. *Electron. Commer. Res. Appl.* 25, 86-94.
16. Kraus, S., Richter, C., Brem, A., Cheng, C.-F., & Chang, M.-L., 2016. Strategies for reward-based crowdfunding campaigns. *J. Innovat. Knowl.* 1(1), 13-23.
17. Kunz, M.M., Bretschneider, U., Erler, M., Leimeister, J.M., 2017. An empirical investigation of signaling in reward-based crowdfunding. *Electronic Commerce Research*, 1-37.

18. LIWC, (Pennebaker Conglomerates) <http://liwc.wpengine.com/>(accessed 1 June 2019).
19. Mitra, T., & Gilbert, E., 2014. The language that gets people to give: Phrases that predict success on kickstarter. CSCW '14: Proc. 17th ACM Conf. Comput. Supported Cooperative Work Soc. Comput. ACM, New York, pp. 49-61.
20. Mollick, E., 2014. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing* 29, 1-16.
21. N. Moy, H. F. Chan and B. Torgler, "How much is too much? The effects of information quantity on crowdfunding performance," *PLOS ONE*, vol. 13, 2018.
22. Parhankangas, A., & Renko, M., 2019. Linguistic style and crowdfunding success among social and commercial entrepreneurs. *J. Bus. Ventur.* 32(2), 215-236.
23. Petruzzelli, A. M., Natalicchio, A., Panniello, U., & Roma, P., 2019. Understanding the crowdfunding phenomenon and its implications for sustainability. *Technological Forecasting and Social Change*, 141, 138-148.
24. Silva L., Silva N.F., Rosa T., 2020. Success prediction of crowdfunding campaigns: a two-phase modeling. *International Journal of Web Information Systems* 16:387–412. doi: 10.1108/ijwis-05-2020-0026.
25. Song Y., Berger R., Yosipof A., Barnes BR., 2019. Mining and investigating the factors influencing crowdfunding success. *Technological Forecasting and Social Change* ,148.
26. Stats. (2019). Retrieved from <https://www.kickstarter.com/help/stats>. (accessed 1.12.2019).
27. Xie, K., Zimei , L., Chen, L., Zhang, W., Liu, S., & Chaudhry, S.S., 2019. Success factors and complex dynamics of crowdfunding: An empirical research on Taobao platform in China. *Electron. Mark.* 29(2), 187-199.
28. Yuan, H., Lau, R.Y.K., & Xu, W., 2016. The determinants of crowdfunding success: A semantic text analytics approach. *Decis. Support Syst.*, 91, 67-76.
29. Zhang, Q., Ye, T., Essaidi, M., Agarwal, S., Liu, V., & Loo, B.T., 2017. Predicting startup crowdfunding success through longitudinal social engagement analysis. *CIKM '17: Proceedings of the 2017 ACM on Conf. Inform. Knowl. Manag.* ACM, New York.

30. Reddy, G., Reddy, M., Lakshmana, K., Kaluri, R., Rajput, D., Srivastava, G., & Baker, T., 2020. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, 8, 54776-54788.
31. Ashokkumar, P., Shankar S. G., Srivastava, G., Maddikunta, P. K., Gadekallu, T. R. (2021). A Two-stage Text Feature Selection Algorithm for Improving Text Classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(3), 1–19.