

Project No. 26-144CS-BS

Identification and Impact of "Sock Puppets" in Wikipedia

זיהוי בובות קש בוויקיפדיה

Final Project Report

Submitted in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

By

Mahmoud Khatib and Daniel Blinov

Supervisor

Dr. Dvir Ross

Department of Computer Science

Sami Shamoon College of Engineering (SCE)

Beer-Sheva

August 2026

תקציר (Hebrew Abstract)

חשבונות "בובות קש" (Sockpuppets) הם זהויות מקוונות מרובות המופעלות על ידי אדם יחיד, לעיתים קרובות במטרה להטות דיונים, ליצור תמיכה מלאכותית, לעקוף חסימות או להשפיע על קונצנזוס בפלטפורמות שיתופיות. זיהוי ידני של חשבונות אלו הוא תהליך גוזל זמן, דורש משאבים רבים ואינו מסוגל להתמודד בקלות עם נפח המידע הגדל, בפרט בוויקיפדיה העברית. השפה העברית מציבה אתגרים ייחודיים לעיבוד שפה טבעית בשל עשירות מורפולוגית, קידומות מחוברות, שונות בכתוב והיעדר ניקוד.

פרויקט זה מציג ארכיטקטורה מולטי-מודאלית אוטומטית לזיהוי בובות קש בוויקיפדיה העברית. המערכת אוספת ומנקה תגובות משתמשים, בונה ייצוגים טקסטואליים ברמת המשתמש ומשלבת שלושה ערוצי ניתוח משלימים: דמיון מילולי המבוסס על TF-IDF Word N-grams, דמיון סגנוני המבוסס על 11 מדדים מבניים, ודמיון סמנטי המבוסס על מודל heBERT המותאם לעברית. שלושת הערוצים מאוחדים באמצעות מסווג Random Forest המנבא האם זוג חשבונות עשוי להיות שייך לאותו אדם.

המודל הוערך באמצעות Leave-One-Out Cross-Validation על 633 זוגות משתמשים, מהם 327 זוגות בובות קש מאומתים ו-306 זוגות לגיטימיים. המודל השיג דיוק כולל של 77%. עבור מחלקת בובות הקש התקבל Precision של 0.81, Recall של 0.72 ו-F1-score של 0.76. ניתוח חשיבות המאפיינים הראה שהרכיב הסמנטי תרם 47.33% לחשיבות החלטת המודל, הרכיב הסגנוני 33.00% והרכיב המילולי 19.67%. בשלב גילוי נוסף הוחל המסווג על זוגות חשבונות שלא תועדו קודם לכן, וזוגות בעלי הסתברות חזויה של 70% ומעלה סומנו לבדיקה אנושית.

Abstract

Sockpuppet accounts are multiple online identities controlled by the same individual, often used to manipulate discussions, create artificial support, evade sanctions, or influence consensus.

Detecting such accounts manually is time-consuming and difficult to scale, especially in Hebrew Wikipedia, where morphological richness, attached prefixes, spelling variation, and the absence of written vowels introduce additional challenges for natural-language processing.

This project presents an automated multimodal framework for detecting sockpuppet accounts in Hebrew Wikipedia. The system collects and preprocesses user comments, constructs user-level textual representations, and combines three complementary detection modalities: TF-IDF word n-gram similarity, stylometric similarity, and semantic similarity based on the Hebrew-trained heBERT transformer. These signals are integrated using a Random Forest classifier that predicts whether two accounts are likely to belong to the same individual.

The framework was evaluated using Leave-One-Out Cross-Validation (LOOCV) on 633 user pairs, including 327 verified sockpuppet pairs and 306 legitimate pairs. The model achieved an overall accuracy of 77%. For the sockpuppet class, precision was 0.81, recall was 0.72, and F1-score was 0.76. Feature-importance analysis showed that semantic features contributed 47.33% of the model's decision importance, stylometry 33.00%, and TF-IDF 19.67%. A subsequent discovery phase applied the trained classifier to previously undocumented account pairs and flagged pairs with predicted probabilities of at least 70% for further investigation.

The expanded report provides additional explanation of the problem formulation, Hebrew-specific preprocessing decisions, the role of each modality, the evaluation protocol, interpretation of the confusion matrix, practical deployment considerations, and threats to validity. These additions are explanatory and do not alter the reported experimental results.

Keywords: Sockpuppet Detection, Hebrew Wikipedia, NLP, heBERT, Stylometry, TF-IDF, Random Forest, Authorship Analysis.

Table of Contents

1. Introduction	6
1.1 Background and Motivation	6
1.2 Problem Definition	7
1.3 Project Objectives	7
1.4 Research Questions	7
1.5 Main Contribution	7
2. Related Work	8
2.1 Stylometric Approaches	8
2.2 Behavioral Analysis	8
2.3 Semantic and Deep-Learning Approaches	8
2.4 Hybrid Models	9
2.5 Motivation for Hebrew NLP Challenges	9
2.6 Positioning of the Proposed Framework	9
3. Dataset and Data Collection	10
3.1 Data Sources	10
3.2 Pair Construction	10
3.3 User-Level Document Construction	10
3.4 Data Quality and Ethical Considerations	11
4. Methodology	12
4.1 End-to-End Pipeline	12
4.2 Hebrew-Specific Preprocessing	12
4.3 TF-IDF Word N-grams	12
4.4 Stylometric Features	13
4.5 Semantic Similarity with heBERT	13
4.6 Multimodal Classification Pipeline	13
4.7 Evaluation Protocol	13
4.8 Evaluation Metrics	14
4.9 Reproducibility Considerations	14
5. Results	15
5.1 Classification Performance	15
5.2 Confusion Matrix Analysis	15
5.3 Feature Importance Distribution	15

5.4 Interpretation of Results	16
6. Discussion.....	17
6.1 Complementarity of Modalities	17
6.2 Practical Meaning of Classification Errors.....	17
6.3 Hebrew-Specific Observations.....	17
6.4 Human-in-the-Loop Use.....	17
7. Discovery Phase.....	18
7.1 Suggested Review Workflow.....	18
8. Limitations and Threats to Validity.....	19
8.1 Threats to Construct Validity	19
8.2 Threats to External Validity.....	19
8.3 Threats to Evaluation Validity.....	19
9. Future Work.....	20
9.1 Proposed Dashboard Components.....	20
10. Conclusions.....	21
11. References.....	22
Appendix A: System Architecture & Data Pipeline Summary	23
Appendix B: Feature Summary.....	24
Appendix C: Evaluation Metrics and Derived Counts	24
Appendix D: Practical Interpretation Guide.....	25

1. Introduction

Online collaborative platforms such as Wikipedia depend on open participation, transparency, and the cumulative work of large communities. This openness makes it possible for contributors with different backgrounds to improve articles, debate content, and reach community decisions. At the same time, the same low barrier to participation can be exploited by users who create and operate multiple accounts.

A sockpuppet is an additional online identity controlled by a person who also controls another account. The existence of multiple accounts is not necessarily harmful in every possible context; the central concern in this project is coordinated use that can distort the apparent number of independent participants, evade restrictions, or influence discussions and consensus. Because the accounts may deliberately appear independent, the detection task is naturally uncertain and requires combining several kinds of evidence.

Manual investigation is difficult to scale. Administrators may need to inspect writing style, editing patterns, discussion history, and contextual clues across many contributions. Such review is valuable because humans can understand context, but it is time-consuming and can be subjective. An automated system is therefore most useful as decision support: it can rank or flag suspicious pairs while leaving the final judgment to human reviewers.

Hebrew Wikipedia creates an additional technical challenge. Hebrew is morphologically rich, many function words appear as prefixes attached to content words, spelling conventions vary, and most text is written without vowel marks. Consequently, two users may express similar ideas with surface forms that are less directly comparable than in simpler lexical settings. This motivates a system that does not depend on only one representation of text.

The proposed project addresses this issue through a multimodal framework. Lexical overlap is modeled using TF-IDF word n-grams, structural writing habits are represented by stylometric measurements, and contextual semantic similarity is captured by heBERT embeddings. A Random Forest classifier combines the three similarity signals and produces a pair-level prediction.

1.1 Background and Motivation

Authorship-related detection problems are rarely solved reliably by a single textual clue. Vocabulary may change with topic, punctuation habits may be consciously modified, and semantic similarity alone may simply indicate that two legitimate editors discuss the same subject. The motivation for multimodal analysis is that different representations fail in different ways. If several partially independent signals agree, the combined evidence can be more informative than any one signal in isolation.

The project is also motivated by the operational setting of Wikipedia. A useful system should not be understood as an automatic banning mechanism. Instead, it should reduce the search space presented to administrators. In this setting, precision is particularly important because a large number of false accusations would reduce trust in the tool. At the same time, recall remains important because missed sockpuppet pairs reduce the value of automated screening.

1.2 Problem Definition

The system receives textual material associated with two user accounts and estimates whether the accounts are likely to be controlled by the same individual. Formally, each pair is represented by three similarity values: lexical similarity, stylometric similarity, and semantic similarity. The classifier maps this compact feature vector to a binary class and a predicted probability.

This is a pairwise classification problem rather than a direct user-identification problem. The system does not attempt to recover a person's real-world identity. It only estimates whether two account-level textual profiles exhibit patterns consistent with the verified sockpuppet examples used during evaluation.

1.3 Project Objectives

1. Dataset Construction: construct an evaluation dataset containing verified sockpuppet pairs and legitimate comparison pairs from Hebrew Wikipedia.
2. Hebrew Preprocessing: develop a preprocessing procedure suitable for noisy Hebrew discussion text and common MediaWiki markup.
3. Feature Extraction: extract complementary lexical, stylometric, and semantic representations from user-generated text.
4. Multimodal Fusion: train a machine-learning classifier that combines the three modalities into a unified prediction.
5. Rigorous Evaluation: evaluate generalization using Leave-One-Out Cross-Validation (LOOCV) and standard classification metrics.
6. Discovery Mechanism: apply the trained framework to previously undocumented user pairs and flag high-probability candidates for human review.

1.4 Research Questions

- Can a compact combination of lexical, structural, and semantic similarity distinguish verified sockpuppet pairs from legitimate pairs in Hebrew Wikipedia?
- Which of the three modalities contributes most strongly to the Random Forest decisions?
- What types of classification errors remain after multimodal fusion?
- Can the trained pipeline be used as a practical candidate-generation mechanism for administrator review?

1.5 Main Contribution

The main contribution is a Hebrew-oriented multimodal detection framework that combines traditional lexical similarity, structural writing-style features, and contextual transformer embeddings. Rather than relying on manually assigned modality weights, the Random Forest learns how to use the three similarity values from labeled pair examples. The project further demonstrates a complete workflow from collection and preprocessing through evaluation and candidate discovery.

2. Related Work

Research on sockpuppet detection intersects with authorship analysis, stylometry, behavioral modeling, semantic representation learning, and anomaly detection. The methods summarized below motivate the design choices of the project while also illustrating why no single signal is sufficient in every case.

2.1 Stylometric Approaches

Stylometry treats writing style as a behavioral fingerprint. Typical measurements include punctuation preferences, sentence and word length, vocabulary richness, character patterns, and syntactic tendencies. Solorio et al. investigated authorship identification on Wikipedia talk pages using stylometric information and showed that recurring stylistic habits can help distinguish authors.

Stylometric evidence has two attractive properties for sockpuppet detection. First, it can remain informative even when two accounts discuss different topics, because structural habits are not identical to content. Second, many such features are inexpensive to compute. Its main weakness is sensitivity to text length and deliberate style modification. Short comments provide fewer observations, and a sophisticated user may consciously alter visible habits.

2.2 Behavioral Analysis

Behavioral approaches focus on how an account acts rather than only on what it writes. Examples include activity timing, editing frequency, interaction networks, repeated participation in the same discussions, and patterns of account creation or sanctions. Kumar et al. studied sockpuppets in online discussion communities and reported recurring behavioral characteristics.

Behavioral signals can complement textual evidence, but they can also be ambiguous. Two legitimate users may be active at similar hours or participate in the same controversial topic. For this reason, the current project focuses on textual modalities while identifying behavioral integration as an important direction for future work.

2.3 Semantic and Deep-Learning Approaches

Lexical methods compare surface words, whereas semantic methods attempt to capture meaning in context. Transformer models such as BERT generate contextual representations in which the vector assigned to a token depends on surrounding text. This makes transformer embeddings useful when two texts express related ideas with different surface wording.

In the present system, heBERT provides a Hebrew-trained semantic representation. Long user documents are divided into sequential chunks, each chunk is embedded, and chunk vectors are aggregated into a user-level representation. Pairwise cosine similarity then measures the semantic proximity of the two user profiles.

Sakib and Spezzano explored BERT-based and behavioral signals for sockpuppet detection. Such work supports the broader idea that contextual representations can contribute information beyond exact word overlap.

2.4 Hybrid Models

Hybrid models combine multiple feature families. Their central assumption is that authorship behavior is multidimensional: vocabulary, structure, semantics, and activity patterns can each provide partial evidence. Mielinytė and Ross evaluated hybrid approaches in an English Wikipedia setting, motivating the multimodal strategy adopted here.

The proposed framework is deliberately compact at the classifier stage. Instead of passing thousands of raw lexical or embedding dimensions directly to the Random Forest, each modality is first converted into a pairwise similarity value. The resulting vector has three interpretable components: Sim_TFIDF, Sim_Stylo, and Sim_heBERT.

2.5 Motivation for Hebrew NLP Challenges

- Morphological complexity: words can vary substantially through inflection by gender, number, tense, and person.
- Attached prefixes: conjunctions, articles, and prepositions may attach directly to a word and create many surface variants.
- Lexical ambiguity: the absence of written vowels can leave the same consonantal form open to multiple interpretations.
- Spelling variation: full and defective spelling conventions can differ across writers.
- Noisy discussion text: Wikipedia talk pages contain markup, links, templates, signatures, and informal writing that must be cleaned before feature extraction.

These characteristics justify a Hebrew-specific preprocessing layer and also strengthen the case for combining lexical and contextual representations. A purely lexical method may fragment semantically related forms, whereas a contextual model can preserve information that is less dependent on exact token identity.

2.6 Positioning of the Proposed Framework

The project positions itself between classical authorship analysis and modern contextual NLP. TF-IDF offers transparency and sensitivity to recurring vocabulary, stylometry captures structural habits, and heBERT captures contextual similarity. The Random Forest serves as the fusion mechanism. The purpose is not to claim that these three signals exhaust all evidence available to administrators, but to test whether their combination provides a useful automated baseline for Hebrew Wikipedia.

3. Dataset and Data Collection

The dataset is built from confirmed sockpuppet cases documented by Wikipedia administrators and a control group of legitimate, highly active editors. User comments are retrieved through the MediaWiki API from discussion-oriented pages such as Talk pages and Noticeboards. Bots are filtered out so that automated text does not contaminate authorship signals.

The unit of evaluation is a pair of users. Verified sockpuppet pairs form the positive class, while legitimate pairs form the control class. The final evaluation dataset contains 633 pairs: 327 verified sockpuppet pairs and 306 legitimate pairs. This yields a nearly balanced class distribution and allows accuracy to remain interpretable alongside class-specific precision, recall, and F1-score.

3.1 Data Sources

The MediaWiki API provides a reproducible interface for collecting public Wikipedia discussion content. Discussion pages are particularly relevant because they contain free-form language and interaction-oriented writing, both of which can reveal stylistic and semantic patterns. The project uses comments rather than relying only on article edits, because article content is often constrained by encyclopedic style and collaborative revision.

3.2 Pair Construction

Pair construction determines the supervised learning examples. Positive examples are based on administrator-confirmed relationships. Negative examples are legitimate comparison pairs. A nearly balanced dataset reduces the risk that a classifier can obtain deceptively high accuracy by predicting only the majority class.

Pair Type	Number of Pairs	Percentage
Legitimate Pairs (Control)	306	48.3%
Verified Sockpuppet Pairs	327	51.7%
Total Evaluation Pairs	633	100.0%

3.3 User-Level Document Construction

Comments belonging to a single user within a specific discussion page are concatenated into a unified user document. This aggregation increases the amount of text available for feature extraction and produces a stable unit for pairwise comparison.

Users with fewer than 10 cleaned characters are excluded. Extremely short samples provide too little evidence for meaningful lexical, stylometric, or semantic comparison. The threshold is intentionally lightweight; it removes unusable cases without claiming that every retained document is equally informative.

3.4 Data Quality and Ethical Considerations

The quality of the labels depends on the administrator-confirmed cases used as ground truth. This is appropriate for supervised evaluation, but it also means the dataset can represent only cases that were detected and documented. Undetected or incorrectly attributed cases are not observable in the ground truth.

The framework operates on account-level public discussion text and predicts relationships between accounts. It should therefore be used as a screening tool rather than as conclusive evidence about a person's identity or intent. The discovery threshold described later is designed to flag candidates for further review, not to automate sanctions.

4. Methodology

4.1 End-to-End Pipeline

7. Collect discussion comments through the MediaWiki API.
8. Filter bots and remove unusable records.
9. Clean MediaWiki markup, URLs, HTML, and special symbols.
10. Construct user-level textual documents.
11. Generate lexical, stylometric, and semantic representations.
12. Convert each representation into a pairwise similarity score.
13. Combine the three scores into $x = [\text{Sim_TFIDF}, \text{Sim_Stylo}, \text{Sim_heBERT}]$.
14. Train and evaluate a Random Forest classifier with LOOCV.
15. Apply the trained pipeline to undocumented pairs and flag high-probability candidates.

The separation between representation and classification is important. Each modality first answers a narrower question—how similar are these two accounts according to a specific view of language? The classifier then learns how the three answers should be combined.

4.2 Hebrew-Specific Preprocessing

Preprocessing removes artifacts that are not intended to represent the author's language while retaining information useful to the three modalities. HTML tags, external URLs, MediaWiki templates, internal-link syntax, headings, and special symbols are removed. Whitespace is normalized by collapsing repeated spaces, tabs, and newlines.

A lightweight Hebrew stemmer strips common prefixes from words longer than four characters. The purpose is to reduce vocabulary fragmentation caused by attached function morphemes without applying aggressive linguistic normalization. This choice reflects a trade-off: more aggressive stemming may merge forms that should remain distinct, whereas no normalization can cause closely related Hebrew forms to appear unrelated to TF-IDF.

Different modalities have different sensitivities to preprocessing. TF-IDF benefits directly from reducing surface variation. Stylometry requires care because punctuation and structural patterns are themselves features. Semantic modeling benefits from removal of markup noise but should preserve enough natural context for heBERT to interpret the text.

4.3 TF-IDF Word N-grams

TF-IDF represents a document according to terms that are frequent in that document but not universally frequent across the corpus. The term-frequency component rewards recurring vocabulary within a user's text, while inverse document frequency reduces the weight of terms that occur in many documents.

The project uses word unigrams and bigrams. Unigrams capture individual lexical choices, while bigrams preserve limited local phrase information. Terms appearing in more than 80% of documents are filtered using $\text{max_df}=0.8$ because extremely common terms provide little discriminative value.

After TF-IDF vectors are constructed, cosine similarity measures the angle between two vectors. A higher cosine similarity indicates that the two users exhibit more similar weighted word and phrase usage. The output of this modality is a single pairwise score, Sim_TFIDF .

4.4 Stylometric Features

Stylometry focuses on how a user writes rather than only on which topics or words appear. The project extracts 11 structural writing metrics, including average word length, average sentence length, word-length distributions, punctuation ratios, and lexical diversity.

Lexical diversity is represented using a Type-Token Ratio normalized to a 100-word sample. Normalization is important because raw vocabulary diversity tends to change with document length. The stylometric features are standardized before comparison so that measurements with larger numerical scales do not dominate the distance calculation.

Pairwise stylometric similarity is derived from Euclidean distance in the standardized feature space. Conceptually, two accounts are stylometrically close when their structural feature profiles are close. This produces the second classifier input, Sim_Stylo.

4.5 Semantic Similarity with heBERT

Semantic similarity is designed to capture contextual resemblance that may not be visible through exact word overlap. The project uses avichr/heBERT, a transformer model pre-trained on Hebrew text.

User documents can be longer than the input size that is convenient for transformer processing. Therefore, each document is split into sequential 120-word chunks. Each chunk is embedded using a mean-pooled representation from the model's final hidden layers. The chunk embeddings are then averaged to produce a user-level semantic vector.

Cosine similarity between two user-level semantic vectors yields Sim_heBERT. Because contextual embeddings encode information beyond exact token identity, this modality can identify similarity even when users employ different surface forms to express related content.

4.6 Multimodal Classification Pipeline

For each user pair, the final feature vector is $x = [\text{Sim_TFIDF}, \text{Sim_Stylo}, \text{Sim_heBERT}]$. This three-dimensional vector is passed to a Random Forest classifier containing 100 decision trees. The classifier estimates $P(\text{Sockpuppet} \mid x)$ and produces a binary prediction.

Random Forest is suitable for this compact fusion task because it can model non-linear interactions and threshold-like relationships among the three similarity values. For example, a pair with moderate lexical similarity but very high semantic and stylometric similarity may be treated differently from a pair with the same average similarity distributed evenly across modalities.

4.7 Evaluation Protocol

The framework is evaluated with Leave-One-Out Cross-Validation (LOOCV) across all 633 labeled pairs. In each iteration, one pair is held out as the test example while the remaining pairs are used for training. The process repeats until every pair has served once as the held-out example.

LOOCV maximizes training-data utilization, which is useful for a dataset of this size. It also produces an out-of-sample prediction for every pair. The aggregated predictions are then used to compute the reported accuracy, precision, recall, F1-score, and confusion matrix.

4.8 Evaluation Metrics

Metric	Interpretation in this project
Accuracy	Fraction of all pairs classified correctly.
Precision (Sockpuppet)	Among pairs predicted as sockpuppets, the fraction that are verified sockpuppet pairs.
Recall (Sockpuppet)	Among verified sockpuppet pairs, the fraction detected by the model.
F1-score	Harmonic balance between precision and recall.
Confusion Matrix	Counts of true negatives, false positives, false negatives, and true positives.

No single metric should be interpreted alone. Accuracy summarizes overall correctness, while precision and recall reveal different operational trade-offs. In an administrator-support setting, false positives create unnecessary investigations, whereas false negatives leave suspicious pairs unflagged.

4.9 Reproducibility Considerations

The pipeline is modular: collection, preprocessing, document construction, feature extraction, classification, and evaluation are conceptually separated. This makes it possible to replace one modality—for example, a newer Hebrew transformer—without redesigning the complete workflow. The report preserves the reported configuration: TF-IDF unigrams/bigrams with max_df=0.8, 11 stylometric metrics, heBERT with sequential 120-word chunks, a 100-tree Random Forest, and LOOCV on 633 pairs.

5. Results

The LOOCV evaluation produced an overall accuracy of 77%. The class-specific results show that the classifier performs meaningfully above a simple chance-level baseline on the nearly balanced dataset, while still leaving substantial room for improvement.

5.1 Classification Performance

Class	Precision	Recall	F1-Score	Support
Not Sockpuppet (0)	0.73	0.82	0.78	306
Sockpuppet (1)	0.81	0.72	0.76	327
Macro Average	0.77	0.77	0.77	633
Weighted Average	0.78	0.77	0.77	633

For the sockpuppet class, precision is 0.81. This means that the model's positive predictions are relatively reliable within the evaluated dataset. Recall is lower at 0.72, indicating that some verified sockpuppet pairs are not detected. The F1-score of 0.76 summarizes this balance.

For the legitimate class, recall is 0.82, showing that most legitimate pairs are correctly rejected as non-sockpuppets. The precision of 0.73 reflects the effect of false negatives from the sockpuppet class entering the set predicted as legitimate.

5.2 Confusion Matrix Analysis

Actual \ Predicted	Legitimate (0)	Sockpuppet (1)
Actual: Legitimate (0)	252 (True Negative)	54 (False Positive)
Actual: Sockpuppet (1)	91 (False Negative)	236 (True Positive)

The model correctly classifies 252 legitimate pairs and 236 sockpuppet pairs. It produces 54 false positives, where legitimate pairs are flagged as sockpuppets, and 91 false negatives, where verified sockpuppet pairs are classified as legitimate.

The asymmetry between 54 false positives and 91 false negatives is consistent with the reported class metrics: sockpuppet precision is stronger than sockpuppet recall. Operationally, this means the classifier is comparatively conservative about positive predictions. When it does predict sockpuppet, it is often correct, but it does not capture every verified case.

5.3 Feature Importance Distribution

Modality	Feature Description	Decision Importance
----------	---------------------	---------------------

Semantic	heBERT contextual embeddings	47.33%
Stylometry	Structural and punctuation metrics	33.00%
TF-IDF	Lexical word n-grams	19.67%

The semantic modality has the largest reported feature importance at 47.33%. This suggests that contextual similarity is the strongest individual signal among the three classifier inputs. Stylometry contributes 33.00%, showing that structural writing habits also provide substantial discriminatory information. TF-IDF contributes 19.67%, indicating that lexical overlap remains useful but is less influential than the two other modalities in the fitted Random Forest.

Feature importance should be interpreted as importance within this specific trained model and feature set. It does not mean that semantic information is universally more important for every sockpuppet case. The value of the multimodal design is precisely that different pairs can be supported by different combinations of evidence.

5.4 Interpretation of Results

The results support the central project hypothesis that combining complementary representations is useful for Hebrew sockpuppet detection. A 77% overall accuracy on a nearly balanced pair dataset indicates that the feature vector contains meaningful information about account relationships.

At the same time, the confusion matrix prevents overstatement. Ninety-one verified positive pairs are missed, and 54 legitimate pairs are incorrectly flagged. The system is therefore not suitable as an autonomous decision-maker. Its appropriate role is prioritization: high-scoring pairs can be sent to administrators together with supporting evidence for further examination.

The relatively strong sockpuppet precision of 0.81 is encouraging for this use case because it limits the proportion of positive predictions that are incorrect. However, the 0.72 recall demonstrates that additional modalities or stronger representations are needed if the goal is to discover a larger fraction of verified cases.

6. Discussion

6.1 Complementarity of Modalities

The three modalities describe different aspects of account behavior. TF-IDF is sensitive to repeated vocabulary and local phrases. Stylometry captures structural habits such as punctuation and length distributions. heBERT captures contextual semantic information. Their combination reduces dependence on any single assumption about how a sockpuppet behaves.

A pair can be lexically dissimilar because the accounts discuss different topics while still showing similar structural habits. Conversely, two legitimate editors can use the same topical vocabulary but differ strongly in style. Semantic similarity can bridge some vocabulary differences, but two users participating in the same discussion can also be semantically similar for legitimate reasons. Multimodal fusion provides a mechanism for balancing these cases.

6.2 Practical Meaning of Classification Errors

A false positive can create unnecessary administrative work and, if interpreted carelessly, can lead to unfair suspicion. Therefore, predictions should be presented as risk indicators rather than accusations. A useful interface would expose the probability score and the component similarities so that reviewers can understand why a pair was flagged.

A false negative is a missed opportunity for automated detection. The 91 false negatives in the evaluation indicate that some verified sockpuppet pairs do not resemble each other strongly enough under the three current signals. Possible explanations include deliberate style changes, insufficient text, topic changes, or behavioral similarities that are not represented in the textual feature vector.

6.3 Hebrew-Specific Observations

The feature-importance distribution is consistent with the motivation for contextual Hebrew modeling. Surface word matching contributes useful information, but Hebrew morphology and spelling variation can weaken exact lexical overlap. The stronger semantic contribution suggests that contextual representation helps recover information that is not captured by word n-grams alone.

Stylometry also remains important despite the language-specific preprocessing. This is useful because structural writing habits are partly independent of Hebrew lexical morphology. The 33.00% importance assigned to stylometry shows that relatively simple, interpretable features can complement a transformer-based representation.

6.4 Human-in-the-Loop Use

The system is best viewed as a human-in-the-loop tool. Automated scoring can reduce the number of account pairs that administrators need to inspect, while humans remain responsible for contextual judgment. This separation is particularly important because similarity does not prove shared control, and platform decisions can have consequences for legitimate contributors.

7. Discovery Phase

Following model validation, the trained pipeline was applied to previously undocumented user pairs extracted from ongoing Wikipedia discussions. The purpose of this phase was candidate generation rather than the creation of new ground-truth labels.

Pairs with a predicted probability $P(\text{Sockpuppet}) \geq 0.70$ were flagged as suspicious candidates for administrative review. The 70% threshold defines a practical screening rule: only pairs with relatively high model confidence are forwarded for further investigation.

The discovery phase should not be interpreted as proof that flagged pairs are sockpuppets. Because the pairs are undocumented, their true status is not established by the model itself. Confirmation requires independent administrative review. This distinction between prediction and verification is essential when deploying authorship-analysis systems in real communities.

7.1 Suggested Review Workflow

16. Run the pair through the preprocessing and three feature-extraction modules.
17. Compute the Random Forest probability.
18. If the probability is below the operational threshold, do not prioritize the pair.
19. If the probability is at least 0.70, present the pair for human review.
20. Expose the three modality scores to support interpretation.
21. Record administrator feedback separately so that future model versions can be evaluated on newly verified data.

8. Limitations and Threats to Validity

Dataset Size: The evaluation contains 633 pairs, which provides a useful baseline but does not cover the full diversity of Hebrew Wikipedia behavior. Larger samples would support more detailed subgroup analysis and more stable estimates.

Ground-Truth Dependency: The labels rely on administrator-verified cases. Sockpuppets that successfully evaded detection are absent from the positive class, and the available cases may overrepresent behaviors that are easier for administrators to identify.

Text-Length Variance: Short comments provide less evidence for vocabulary, style, and semantic profiling. Although extremely short cleaned texts are excluded, the retained documents can still vary substantially in length.

Lightweight Stemmer: The rule-based Hebrew prefix stemmer does not perform full morphological disambiguation. It may fail to normalize some related forms or may remove a prefix in a context where that operation is not linguistically ideal.

Topic Effects: Users discussing the same subject can share vocabulary and semantics for legitimate reasons. Conversely, one person using multiple accounts across different topics may show weaker lexical and semantic similarity.

Pairwise Framing: The classifier evaluates pairs independently. It does not directly model larger groups of accounts or transitive relationships among multiple suspected identities.

Model Interpretation: Random Forest feature importance summarizes contribution at the model level but does not provide a complete causal explanation for individual predictions.

Generalizability: The system is designed and evaluated for Hebrew Wikipedia. Performance should not be assumed to transfer unchanged to other languages, platforms, or community norms.

8.1 Threats to Construct Validity

The project operationalizes sockpuppet similarity through three textual modalities. This is a practical representation, but sockpuppetry is a broader behavioral phenomenon. A user may intentionally alter language while retaining non-textual behavioral patterns. Therefore, the model measures textual evidence associated with verified pairs rather than the complete concept of coordinated account control.

8.2 Threats to External Validity

The data source, language, and community context constrain external validity. Hebrew Wikipedia has its own conventions and contributor population. Applying the same model to another language or platform would require new preprocessing choices, new training data, and independent evaluation.

8.3 Threats to Evaluation Validity

LOOCV produces a prediction for every pair and makes efficient use of the available data. However, the report's current evaluation is pair-based. If the same individual contributes to multiple related pairs, future work should examine group-aware splitting strategies to measure whether performance changes when identities rather than pairs are separated between training and test data. This is a proposed validation extension; it does not alter the reported LOOCV results.

9. Future Work

Advanced Hebrew Morphology: Integrate deeper Hebrew morphological analyzers such as Dicta or YAP to improve token normalization and linguistic disambiguation.

Behavioral Signals: Add edit timestamps, revision histories, interaction patterns, and other non-textual behavior. These signals may help cases where a user intentionally changes writing style.

Newer Hebrew Transformers: Evaluate newer Hebrew language models such as AlephBERT or DictaBERT and compare their semantic representations with heBERT.

Interactive Dashboard: Build a web dashboard for administrators that displays ranked suspicious pairs, prediction probabilities, component similarities, and review status.

Threshold Calibration: Study how different decision thresholds change precision and recall so that the system can be tuned for conservative screening or broader discovery.

Explainability: Add per-pair explanations that show which modality most strongly supports a prediction and provide non-sensitive summary evidence for reviewers.

Group Detection: Extend pairwise classification into graph-based clustering so that multiple accounts suspected of belonging to the same operator can be analyzed jointly.

Larger and Temporal Evaluation: Collect additional verified cases and evaluate the system on later time periods to test robustness as language and community behavior evolve.

9.1 Proposed Dashboard Components

Component	Purpose
Candidate Queue	Rank pairs by predicted sockpuppet probability.
Modality Panel	Show TF-IDF, stylometric, and heBERT similarity values.
Review Status	Track pending, reviewed, confirmed, or rejected candidates.
Evidence Links	Provide links to relevant public discussion history for manual examination.
Model Version	Record which model configuration produced each score.

10. Conclusions

This project establishes an automated multimodal detection pipeline tailored to Hebrew Wikipedia. The framework combines TF-IDF word n-grams, 11 stylometric metrics, and heBERT semantic embeddings, then fuses the three pairwise similarity signals using a 100-tree Random Forest classifier.

Evaluation with Leave-One-Out Cross-Validation on 633 pairs produced 77% overall accuracy. For the sockpuppet class, precision reached 0.81, recall 0.72, and F1-score 0.76. The confusion matrix contains 236 true positives, 252 true negatives, 54 false positives, and 91 false negatives.

Feature importance shows that heBERT semantic similarity contributes 47.33% of the model's decision importance, stylometry contributes 33.00%, and TF-IDF contributes 19.67%. This distribution supports the multimodal design: contextual semantics provides the strongest individual signal, while structural and lexical information remain meaningful complementary sources.

The discovery phase demonstrates how the model can be used beyond retrospective evaluation. Previously undocumented pairs with predicted probability of at least 70% are flagged for administrator review. The appropriate interpretation is decision support, not automatic enforcement.

Overall, the project provides a practical Hebrew-oriented baseline and a modular architecture for future development. The most promising extensions are deeper Hebrew morphology, behavioral signals, newer transformer models, larger evaluation datasets, group-aware validation, and an interactive review dashboard.

11. References

1. Kumar, S., Spezzano, F., Subrahmanian, V. S., & Faloutsos, C. (2017). Banned on the web: Detecting sockpuppets in online discussion forums. *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 1271–1280.
2. Mielinytė, A., & Ross, D. (2022). Hybrid models for sockpuppet account detection in online collaborative environments. *Journal of Information Security and Applications*, 68, 103210.
3. Sakib, N., & Spezzano, F. (2020). Sockpuppet detection in Wikipedia using BERT embeddings and behavioral signatures. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 415–422.
4. Solorio, T., Ragusha, S., Mendenhall, M., & Montes-y-Gómez, M. (2013). Authorship identification on Wikipedia talk pages using stylometric features. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, 412–424.

Appendix A: System Architecture & Data Pipeline Summary

Stage	Input	Main Operation	Output
Data Acquisition	MediaWiki discussion pages	Retrieve comments and filter bots	Raw user comments
Preprocessing	Raw comments	Remove markup/URLs; normalize whitespace; Hebrew prefix stemming	Clean text
Document Construction	Clean comments	Concatenate comments by user and discussion page	User-level documents
Lexical Features	User documents	TF-IDF unigram/bigram vectors + cosine similarity	Sim_TFIDF
Stylometric Features	User documents	11 standardized structural metrics + distance-based similarity	Sim_Stylo
Semantic Features	120-word chunks	heBERT embedding, mean pooling, cosine similarity	Sim_heBERT
Classification	Three similarity scores	Random Forest with 100 trees	Class + probability
Evaluation	633 labeled pairs	LOOCV	Accuracy, precision, recall, F1, confusion matrix
Discovery	Undocumented pairs	Apply trained model and threshold	Candidates with $P \geq 0.70$

Appendix B: Feature Summary

Feature Family	Examples / Configuration	Role
TF-IDF	Word unigrams and bigrams; max_df=0.8	Capture lexical overlap and recurring phrase usage.
Stylometry	11 metrics including word/sentence length, punctuation ratios, word-length distribution, normalized TTR	Capture structural writing habits.
heBERT	Sequential 120-word chunks; mean-pooled hidden representations; user-vector averaging	Capture contextual semantic similarity.
Fusion	$x = [\text{Sim_TFIDF}, \text{Sim_Stylo}, \text{Sim_heBERT}]$	Provide compact multimodal input to classifier.

Appendix C: Evaluation Metrics and Derived Counts

The following formulas explain the standard metrics used in the report. They are included for clarity; the reported values remain those obtained by the project evaluation.

Metric	Formula / Meaning
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$
Precision	$TP / (TP + FP)$
Recall	$TP / (TP + FN)$
F1-score	$2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$

Using the reported confusion matrix: TP = 236, TN = 252, FP = 54, and FN = 91. These counts sum to 633 evaluated pairs and are consistent with the reported overall accuracy after rounding.

Appendix D: Practical Interpretation Guide

A high model probability should be treated as a prioritization signal. The model does not observe real-world identity and cannot independently establish intent. Reviewers should combine model output with platform-specific evidence and administrative procedures.

- High TF-IDF similarity: the accounts use similar weighted vocabulary or short phrases.
- High stylometric similarity: the accounts exhibit similar structural writing patterns.
- High heBERT similarity: the account-level texts are contextually close in semantic representation.
- High combined probability: the Random Forest has learned that the observed combination resembles positive examples in the labeled dataset.
- Probability ≥ 0.70 in the discovery phase: candidate for human review, not automatic confirmation.