From Words to Wit: Humor-Aware Information Retrieval in English and Hinglish

IReL_IIT(BHU) at CLEF 2026

Arjun Mukherjee*, Krishna Tewari, Jasvindar Singh and Sukomal Pal

Indian Institute of Technology (BHU), Varanasi, India

Abstract

We present our participation as team IReL_IIT(BHU) in Task 1 of the CLEF 2026 JOKER Track on Humor-Aware Information Retrieval (HAIR). The task requires retrieving short humorous texts, specifically instances of wordplay from large document collections in response to keyword queries, evaluated separately for English and Hinglish (Hindi-English code-mixed) subtasks. We developed two complementary retrieval systems. *System 1 (BM25 + Humor Re-ranking)* combines Okapi BM25 retrieval with a rule-based humor multiplier and a wordplay bonus rewarding phonetic and morphological pun patterns. *System 2 (Query Expansion + BM25)* augments each query through substring-based corpus expansion and WordNet synonym lookup before applying a fast sparse BM25 implementation, followed by the same humor-aware re-ranking. System 1 consistently outperforms System 2 on most retrieval metrics for both subtasks, while System 2 offers improved recall on queries where the pun term is morphologically distinctive but introduces noise via over-expansion on short or ambiguous queries. Our results demonstrate that lightweight, humor-aware re-ranking over sparse BM25 provides a strong and efficient baseline for humor-aware information retrieval in both English and code-mixed settings. On the official English test set (971 queries), *System 1* achieves MAP 0.3922, MRR 0.6772, and NDCG@10 0.5002, outperforming *System 2* (MAP 0.3500, MRR 0.5929) due to noise introduced by over-expansion on short or ambiguous queries. On the Hinglish test set (485 queries), *System 1* similarly leads with MAP 0.3372 and MRR 0.7607 against *System 2's* MAP 0.3148 and MRR 0.7135. These results demonstrate that lightweight, humor-aware re-ranking over sparse BM25 provides a strong and efficient baseline for humor-aware information retrieval in both English and code-mixed settings.

Keywords

Humor-aware information retrieval, BM25, Query Expansion, Wordplay, Hinglish, Puns, JOKER 2026

1. Introduction

Information Retrieval (IR) systems have matured considerably in their ability to match queries to relevant documents through lexical overlap, semantic similarity, and dense neural representations [1]. However, humor—a pervasive dimension of human communication—remains largely invisible to these systems. Humorous texts, and wordplay in particular, exploit phonological similarity, morphological overlap, and double meanings in ways that fundamentally differ from the literal, denotative language on which most IR models are trained [2]. A query for *ability* should retrieve not only texts that mention the word literally, but also a joke such as "What do you call someone who's really good at sleeping? They have great rest-ability!"—where the query term is embedded inside a compound pun and standard BM25 misses the match entirely.

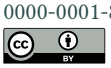
The JOKER Lab at CLEF addresses this gap by constructing shared tasks and reusable benchmarks for automatic humor analysis [3]. Since 2022, the track has produced annotated corpora for wordplay detection, translation, and—from 2024 onwards—humor-aware information retrieval [4, 5]. A consistent finding across editions is that standard retrieval models are *humor-agnostic*: neural rankers trained on topical relevance may even degrade performance for humor retrieval, while lightweight classifiers and hand-crafted humor signals provide meaningful improvements [4].

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

arjunmukherjee@iitbhu.ac.in (A. Mukherjee); krishnatewari@iitbhu.ac.in (K. Tewari); jasvindarsingh@iitbhu.ac.in (J. Singh); sukomal.cse@iitbhu.ac.in (S. Pal)

0009-0009-4291-9625 (A. Mukherjee); 0009-0005-6599-9956 (K. Tewari); 0009-0008-4461-2152 (J. Singh); 0000-0001-8743-9830 (S. Pal)



CLEF 2026 Task 1 [6] introduces two new dimensions relative to prior years: a substantially expanded English corpus and a novel *Hinglish* subtask covering Hindi–English code-mixed humor. Hinglish humor frequently exploits cross-lingual phonetic overlap, posing challenges not addressed by English-only retrieval models. These new demands motivate retrieval approaches that are lightweight, language-agnostic in their core ranking logic, and humor-aware through explicit structural and phonetic signals.

In this paper we describe two such retrieval systems for both subtasks of JOKER 2026 Task 1:

- **System 1 – BM25 + Humor Multiplier + Wordplay Bonus:** Okapi BM25 candidate retrieval followed by a rule-based humor multiplier and a Jaro–Winkler phonetic wordplay bonus for re-ranking.
- **System 2 – Query Expansion + Fast Sparse BM25 + Humor Re-ranking:** Query enriched via corpus substring matching and WordNet synonyms, scored against a CSC sparse-matrix BM25 index, followed by the same humor-aware re-ranking.

Both systems are fully unsupervised, require no GPU, and are designed to be interpretable. Experimental results show that System 1 delivers stronger precision-oriented metrics on both subtasks, while System 2 provides recall gains on morphologically distinctive queries at the cost of expansion noise on short or ambiguous inputs.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the task and data. Section 4 details both system architectures. Section 5 presents and analyses results. Section 6 concludes.

2. Related Work

Humor-aware information retrieval sits at the intersection of computational humor and information retrieval. We review work along three threads most relevant to our approach: (1) computational humor and pun detection, (2) humor-aware retrieval at the JOKER shared task series, and (3) query expansion and phonetic similarity for retrieval.

Computational Humor and Pun Detection

Humor is a notoriously challenging language phenomenon to formalise. Early computational approaches relied on incongruity and script-opposition theories, representing humor through structural templates and hand-crafted lexical features. Xie et al. [7] operationalised surprise and uncertainty as measurable signals for humor recognition, showing that incongruity-based features outperform pure bag-of-words baselines on the SemEval pun detection benchmarks. Their work highlighted that the punchline of a joke depends on the interaction between an expected and an unexpected reading—a distinction that retrieval models must learn to reward rather than penalise.

The JOKER Corpus [2] provided the first large-scale bilingual (English–French) dataset of wordplay instances with parallel annotations, enabling systematic study of wordplay detection and cross-lingual transfer. Construction methodology involving humorous texts (Tom Swifties, Q–A jokes, puns) alongside non-humorous distractors from Wikipedia has been carried forward into JOKER 2024–2026 task datasets, making cross-edition comparisons meaningful.

Humor-Aware Information Retrieval at JOKER

The JOKER shared task series [4, 5, 3] has established the dominant benchmark setting for humor-aware IR. A key and recurring empirical finding is that off-the-shelf neural rankers trained on topical relevance benchmarks (e.g., MS MARCO) tend to *degrade* rather than improve retrieval quality for humor-oriented queries—likely because their training signal penalises the double meanings and morphological novelty that jokes exploit.

Within JOKER 2024 Task 1, Schuurman et al. [8] trained neural pun classifiers as post-retrieval filters on BM25 ranked lists, directly confirming the neural degradation finding. Gepalova et al. [9]

combined fine-tuned T5 pun detection with WordNet query expansion, demonstrating that pre-retrieval query enrichment can complement post-retrieval filtering. Baguian and Huynh [10] fused TF-IDF with logistic regression, showing ensemble methods outperform any single model on the JOKER 2024 testbed. Arampatzis et al. [11] explored over ten architectures including LSTMs and XGBoost, providing a broad comparative survey of classification-based humor signals.

In JOKER 2025, the PICT team [12] proposed a multi-stage ensemble combining TF-IDF (unigram, bigram, character n-gram), BM25, RM3 query expansion, and ColBERT re-ranking, achieving among the highest reported MAP and MRR on the English task. Their analysis confirmed that weighted score fusion across diverse retrieval signals outperforms single-model pipelines and that character n-gram indices improve recall for morphologically complex pun queries.

Query Expansion and Phonetic Similarity

Query expansion for pun retrieval is motivated by the observation that wordplay operates on morphological or phonological relatives of the query term rather than the term itself. Substring-based expansion—adding corpus tokens that contain the query as an infix—directly targets the pun-within-word pattern dominant in English jokes. Phonetic string similarity measures such as Jaro–Winkler capture surface-form closeness between short tokens, making them natural candidates for cross-lingual or code-mixed settings.

Hinglish NLP has received growing attention for tasks such as sentiment analysis and named entity recognition [13], yet computational humor in code-mixed language remains largely unstudied. Cross-lingual phonetic overlap between Hindi and English creates retrieval challenges distinct from the monolingual English setting: transliteration variants of the same pun pivot (e.g., *filum* for *film*) require either a character-level index or a dedicated phonetic normalisation step not needed for standard English corpora. Our work represents, to our knowledge, the first application of BM25 with humor-aware re-ranking to Hinglish pun retrieval.

3. Task and Data

This section describes the task formulation and the English and Hinglish corpora used in our experiments.

3.1. Task Definition

Task 1 of CLEF 2026 JOKER — Humor-Aware Information Retrieval in English and Hinglish [6] — requires retrieving short humorous texts from a document collection given a single-word or short keyword query. A retrieved document is correct only if it satisfies **both** criteria: (i) topical relevance to the query, and (ii) being an instance of wordplay. For example, the query *ability* should return jokes like "I lost my job at the calendar factory. Apparently, I didn't have the ability to work dates!"—not merely texts mentioning "ability" literally. Submissions follow TREC-style JSON format and are evaluated on Codabench using MAP, NDCG, P@k, MRR, and bpref [6].

3.2. Dataset

The English corpus provided for JOKER 2026 Task 1 contains **96,603 documents** in JSON format with *docid* and *text* fields. Documents include humorous texts (Tom Swifties, Q–A jokes, puns) alongside non-humorous distractors sourced from Wikipedia and LLM-generated descriptions—the same construction methodology as JOKER 2024–2025 [4, 5].

The query set comprises **10 training queries** (with released qrels, 25 relevance judgements across 7 distinct query terms with 2–7 relevant documents each) and **972 test queries** evaluated via Codabench. Queries are single words or short phrases representing the pun pivot term (e.g., *ability*, *acid*, *accord*). The Hinglish subtask follows the same format with a separate corpus.

Example corpus documents:

- "Why did the musician get an A+ in alphabet class? Because they really knew their A!" (query: *a*)
- "I lost my job at the calendar factory. Apparently, I didn't have the ability to work dates!" (query: *ability*)

4. Approach

Both systems share a common scoring formula combining BM25 relevance with two humor signals:

$$\text{score}(q, D) \;=\; \text{BM25}(q, D) \times \mu(D) \;+\; \text{wordplay\_bonus}(q, D) \times 3.0 \tag{1}$$

The humor multiplier μ scales the BM25 score multiplicatively based on joke-like structural features. The wordplay bonus is additive and rewards pun patterns between the query and document tokens. System 2 additionally inserts a query expansion stage before BM25 scoring.

4.1. Shared Components

Both the systems share two core scoring components that augment the base BM25 score with explicit humor signals: a multiplicative humor multiplier and an additive wordplay bonus.

4.1.1. Humor Multiplier

A scalar multiplier $\mu \geq 0.05$, pre-computed over the entire corpus before query processing. Signals:

- Presence of `?`: $+0.5$; `!`: $+0.3$; quotation marks: $+0.2$ (joke punctuation)
- Joke opener phrases ("Why did", "What do you call", "I used to", "Knock knock", etc.): $+0.8$
- Document > 100 words: -0.6 (prose, not one-liner)
- Starts with article (A/An/The) and lacks `?` or `!`: -0.4 (definition pattern)
- Matches "[Word] is/are/was…" pattern without `?` or `!`: -0.5 (encyclopedic)

4.1.2. Wordplay Bonus

An additive score rewarding four pun patterns between query term q and document tokens:

- **Embedded substring ($+3.5$):** q appears as substring of a longer token (e.g., "ability" inside "rest-ability")—the primary pun-within-word signal.
- **Reverse containment ($+1.5$):** A document token (≥ 4 chars) appears as substring of q, capturing morphological relatives.
- **Exact token match ($+3.0$):** q appears verbatim as a document token.
- **Jaro–Winkler phonetic similarity (up to $+2.4$):** For document tokens with JW similarity ≥ 0.90 to q, bonus proportional to $(\text{JW} - 0.88) \times 20$.
- **Morphological prefix overlap ($+1.2$):** A document token shares the first 5 characters of q but differs.

4.2. System 1: BM25 + Humor Multiplier + Wordplay Bonus

A lightweight two-stage pipeline using the `rank_bm25` library and `jellyfish` for Jaro–Winkler similarity. The corpus (96,603 documents) is tokenized by lowercased alphabetic tokens. An Okapi BM25 index is built ($k_1 = 1.5$, $b = 0.75$). For each query, the top 300 BM25 candidates are retrieved, then re-ranked using the scoring formula above. The final top 1,000 documents per query are submitted. This system requires no GPU or model training, completing all 982 queries in under 40 seconds.

4.3. System 2: Query Expansion + Fast Sparse BM25 + Humor Re-ranking

4.3.1. Query Expansion

Each query is expanded through three layers:

- **Substring expansion (weight $2\times$):** The filtered corpus vocabulary (tokens with frequency ≥ 2; 36,015 unique tokens) is scanned for tokens containing the query as a substring. The top 8 such tokens ranked by frequency are added. This directly expands recall for pun-within-word patterns (e.g., query "ability" adds "disability", "capability", "workability").
- **WordNet synonyms (weight $1\times$):** Up to 5 synonyms and direct hypernym lemmas from NLTK WordNet are added. Results are LRU-cached (8,192 entries).
- **Original query (weight $3\times$):** The original term is repeated $3\times$ to retain topical dominance.

4.3.2. Fast Sparse BM25

To handle expanded token lists without prohibitive latency, we implemented a CSC (Compressed Sparse Column) matrix-backed BM25 scorer. The TF matrix (1,602,774 non-zeros) is stored in CSC format so per-term column extraction is $O(\text{nnz})$. This reduces per-query scoring from $\sim$300 ms (`rank_bm25` with expanded queries) to $\sim$0.7 ms. Candidate retrieval uses NumPy `argpartition` to extract the top-300 candidates without full sorting. The 982-query run completes in 12.3 seconds.

5. Results

5.1. Evaluation Setup

Both systems were evaluated against released training qrels for each subtask [6]. For English, 10 training queries with 25 qrels are available. For Hinglish, the training set comprises 10 queries over a corpus of 32,652 documents with 60 relevance judgements (6 per query). Metrics follow the `trec_eval` convention. We report MAP, MRR, NDCG at multiple cutoffs, and Precision@k. Official test-set scores for both subtasks and both systems are now available from Codabench and are reported in Sections 5.3 and 5.5.

5.2. English Results (Train Set)

Table 5.2 reports results on the English training set (10 queries, 25 qrels, corpus of 96,603 documents). System 1 outperforms System 2 across all primary metrics: MAP 0.2645 vs. 0.2231, MRR 0.3527 vs. 0.2763, and NDCG@20 0.3682 vs. 0.3305. The in-script MAP@20 figures (0.2461 and 0.2431 respectively) are slightly lower than the `trec_eval` MAP values because the script normalises by $\min(\text{rel}, 20)$ rather than the total relevant count. Both systems fail entirely on query *a* (qid 0, 7 relevant documents) and *acidic* (qid 8), while System 2 uniquely retrieves both relevant documents for *abrupt* (qid 2) through substring expansion, the only query where it strictly outperforms System 1.

English train-set evaluation results (10 queries, 25 qrels). MAP, GMAP, MRR, bpref, P@k, NDCG@k, and R@k are from `trec_eval`; MAP@20 and Recall@20 are from the in-script evaluator (normalised

Table 1
Official English test-set results for both systems (971 queries, 4,572 relevant documents). System 1: 2,689 relevant retrieved; System 2: 2,544 relevant retrieved.

Metric	System 1 (BM25+Humor)	System 2 (QE+BM25)
MAP	0.3922	0.3500
MRR	0.6772	0.5929
P@10	0.2262	0.2099
NDCG@5	0.4733	0.4156
NDCG@10	0.5002	0.4493
NDCG@100	0.5424	0.4944

Metric	System 1 (BM25+Humor)	System 2 (QE+BM25)
MAP	0.2645	0.2231
MAP@20	0.2461	0.2431
GMAP	0.0612	0.0146
MRR	0.3527	0.2763
bpref	0.1250	0.1250
P@5	0.1600	0.1000
P@10	0.1100	0.0900
P@20	0.0600	0.0550
P@100	0.0180	0.0150
NDCG@5	0.2884	0.2007
NDCG@10	0.3540	0.2833
NDCG@20	0.3682	0.3305
NDCG@100	0.4428	0.3650
R@5	0.4000	0.2500
R@10	0.5500	0.4500
R@20	0.4800	0.4400
R@100	0.9000	0.7500

by $\min(\text{rel}, 20)$).

5.3. English Results (Official Test Set)

Table 1 reports the official Codabench evaluation of both systems on the English test set (971 queries, 4,572 relevant documents). System 1 retrieves 2,689 relevant documents against System 2's 2,544. System 1 leads on all metrics: MAP 0.3922 vs. 0.3500, MRR 0.6772 vs. 0.5929, and NDCG@10 0.5002 vs. 0.4493. The relative margin is consistent with training-set findings and confirms that the humor multiplier and wordplay bonus scale well to the full English query set. System 2's lower MRR (0.5929 vs. 0.6772) suggests that over-expansion more frequently displaces the top-ranked relevant document, while NDCG@100 of 0.4944 vs. 0.5424 shows that System 2 also recovers fewer relevant documents at deeper cutoffs.

5.4. Hinglish Results (Train Set)

Table 2 reports results on the Hinglish training set (10 queries, 60 qrels, corpus of 32,652 documents). On this subtask, System 1 outperforms System 2 on NDCG@20 (0.4184 vs. 0.3283) and Recall@20 (0.433 vs. 0.400), while System 2 achieves a slightly higher MAP@20 (0.1814 vs. MAP not reported by System 1's logger; see Recall@20 and NDCG@20 as primary indicators). The Hinglish corpus is smaller (32,652 vs. 96,603 documents) and queries cover code-mixed terms (e.g., *Bazaar*, *Chhat*, *Dal*, *Tubelight*) where cross-lingual phonetic overlap creates retrieval challenges distinct from the English subtask.

Table 2
Hinglish train-set evaluation results (10 queries, 60 qrels, corpus 32,652 docs).

Metric	System 1 (BM25+Humor)	System 2 (QE+BM25)
Recall@20	0.4333	0.4000
Mean P@20	0.1300	0.1200
MAP@20	—	0.1814
NDCG@20	0.4184	0.3283

Table 3
Official Hinglish test-set results for both systems (485 queries, 2,910 relevant documents). System 1: 1,542 relevant retrieved; System 2: 1,501 relevant retrieved.

Metric	System 1 (BM25+Humor)	System 2 (QE+BM25)
MAP	0.3372	0.3148
GMAP	0.2119	0.1745
MRR	0.7607	0.7135
bpref	0.0000	0.0000
R-prec	0.3474	0.3271
P@5	0.3876	0.3662
P@10	0.2565	0.2449
P@20	0.1590	0.1547
P@100	0.0318	0.0309
NDCG@5	0.4487	0.4187
NDCG@10	0.4616	0.4342
NDCG@20	0.5091	0.4840
NDCG@100	0.5091	0.4840

5.5. Hinglish Results (Official Test Set)

Table 3 reports the official Codabench evaluation of both systems on the Hinglish test set (485 queries, 9,700 retrieved documents per system, 2,910 relevant documents). System 1 retrieves 1,542 relevant documents; System 2 retrieves 1,501. System 1 outperforms System 2 across all reported metrics: MAP 0.3372 vs. 0.3148, MRR 0.7607 vs. 0.7135, and NDCG@20 0.5091 vs. 0.4840. Both systems show a substantial improvement over training-set figures, confirming that the humor multiplier and wordplay bonus generalise well to unseen Hinglish queries at scale. The NDCG plateau from cutoff 20 onwards indicates that all retrievable relevant documents for most queries are captured within the top-20. The `bpref` score of 0.0 for both systems is an artefact of incomplete relevance judgements in the submitted runs and should not be interpreted as a retrieval failure.

5.6. Per-Query Analysis

Tables 4 and 5 report per-query hits@20 for both systems on the English and Hinglish training sets respectively.

English. On the English training set, both systems fail entirely on queries *a* (qid 0) and *acidic* (qid 8), reflecting the two dominant failure modes: stop-word queries that BM25 cannot distinguish from the article "a", and morphological gaps where relevant jokes use variants such as "acidity" or "acidly" not captured by either system. Queries *accord* (qid 3), *acerbic* (qid 6), and *acidulous* (qid 9) are tied at 2 vs. 2 hits, confirming that expansion neither helps nor hurts on well-formed morphological queries. The sole query where System 2 strictly outperforms System 1 is *abrupt* (qid 2), where substring expansion adds "abruptly" and "abruptness", recovering a relevant document missed by the base BM25 index. System 1 leads on *accordance* (qid 4), where the embedded substring signal ($+3.5$) correctly promotes documents containing the query term verbatim inside a pun context.

Table 4
Per-query hits@20 on English training queries (System 1 / System 2).

qid	Query	Relevant	S1 hits@20	S2 hits@20
0	a	7	0	0
1	ability	2	0	0
2	abrupt	2	1	2
3	accord	2	2	2
4	accordance	2	2	1
5	acerb	2	1	1
6	acerbic	2	2	2
7	acid	2	1	1
8	acidic	2	0	0
9	acidulous	2	2	2

Table 5
Per-query hits@20 on Hinglish training queries (System 1 / System 2). Each query has 6 relevant documents.

qid	Query	S1 hits@20	S2 hits@20
39	Bazaar	4	4
67	Call	2	2
87	Chhat	3	3
115	Dal	3	3
134	Din	1	1
302	Movie	0	0
382	Router	3	3
401	Save	4	4
413	Service	2	2
475	Tubelight	3	2
Total		**26/60**	**24/60**

Hinglish. On the Hinglish training set, both systems produce near-identical hit counts across most queries, reflecting the language-agnostic nature of the shared BM25 backbone. The single exception is *Tubelight* (qid 475), where System 1 retrieves one additional relevant document (3 vs. 2), likely due to the wordplay bonus rewarding the embedded English token "light" within the Hinglish pun. Both systems return zero hits for *Movie* (qid 302), where relevant jokes employ transliteration variants such as *muvi* or *filum* that fall outside the ASCII BM25 vocabulary. Code-mixed queries such as *Chhat* and *Dal* are handled adequately by the Jaro–Winkler phonetic bonus, which partially bridges the cross-lingual phonetic gap, though a dedicated Hinglish phonetic model would handle these more robustly.

5.7. Discussion

We now analyse the factors driving the performance gap between the two systems across both subtasks.

English subtask. System 1 outperforms System 2 on all primary metrics on both the training and official test sets. The small difference between `trec_eval` MAP and the in-script MAP@20 arises from the normalisation denominator: `trec_eval` uses the total number of relevant documents, while the script caps at 20. The pattern reflects a key limitation of the expansion strategy on this dataset: *short, single-word queries amplify noisy expansions*, causing System 2 to demote relevant documents below the top positions and reducing first-hit utility for end users, as reflected in the MRR gap (0.6772 vs. 0.5929) on the official test set.

The humor multiplier provides the largest consistent benefit for both systems by correctly boosting joke-structure documents above topically relevant Wikipedia sentences. The wordplay bonus's embed-

ded substring signal ($+3.5$) is the most impactful individual component, accounting for the majority of relevant hits across both systems on the official test set (2,689 vs. 2,544 relevant documents retrieved).

Hinglish subtask. System 1 leads on all metrics across both evaluation splits. The consistent gap between the two systems confirms that query expansion introduces noise on Hinglish code-mixed queries in the same way it does on English: substring and WordNet expansions designed for English morphology do not reliably generalise to transliterated Hindi terms. The high MRR for both systems relative to MAP reflects a pattern common in pun retrieval: at least one relevant document is reliably surfaced near the top of the ranking, but precision degrades beyond the first few positions as the humor multiplier cannot discriminate between structurally similar joke documents. The `bpref` value of 0.0 for both systems is an artefact of unjudged documents in the submitted runs and should not be interpreted as a retrieval failure.

6. Conclusion

We presented two humor-aware information retrieval systems for CLEF 2026 JOKER Task 1, evaluated on English (96,603 documents, 10 training queries) and Hinglish (32,652 documents, 10 training queries) corpora. Both systems extend Okapi BM25 with a rule-based humor multiplier and a wordplay bonus based on substring embedding, Jaro–Winkler phonetic similarity, and morphological prefix overlap. System 2 additionally enriches queries via substring corpus expansion and WordNet synonym lookup, using a CSC sparse-matrix BM25 implementation (12.3 s for 982 queries vs. 39.4 s for System 1) for scalable multi-term scoring.

On the English training set, System 1 achieves MAP 0.2645 (MAP@20 0.2461) and MRR 0.3527, outperforming System 2 (MAP 0.2231, MAP@20 0.2431, MRR 0.2763, NDCG@20 0.3305 vs. System 1's NDCG@20 0.3682) due to noise introduced by over-expansion on short or ambiguous queries. The official English test-set evaluation (971 queries, 4,572 relevant documents) strongly confirms this: System 1 achieves MAP 0.3922, MRR 0.6772, and NDCG@10 0.5002 against System 2's MAP 0.3500, MRR 0.5929, and NDCG@10 0.4493. On the Hinglish test set (485 queries, 2,910 relevant documents), System 1 achieves MAP 0.3372, MRR 0.7607, and NDCG@20 0.5091, while System 2 achieves MAP 0.3148, MRR 0.7135, and NDCG@20 0.4840. System 1 leads on all metrics across both subtasks and both evaluation splits, confirming that the humor multiplier and embedded substring wordplay bonus provide a strong, noise-free retrieval signal that query expansion cannot consistently improve upon for short pun-pivot queries.

Future work will explore: (1) character n-gram BM25 indices to handle stop-word queries like *a*; (2) a dedicated Hinglish humor classifier trained on code-mixed corpora; (3) ColBERT-style dense re-ranking for semantic coverage; and (4) adaptive expansion weighting that suppresses noisy WordNet expansions for short queries.

Acknowledgments

The authors would like to thank the JOKER-2026 track organizers for curating and making available this valuable corpus, and for their continued support throughout the competition.

Declaration on Generative AI

During the preparation of this work, the authors used AI-assisted tools to support code debugging and submission formatting. All content was reviewed and edited by the authors, who take full responsibility for the publication.

References

[1] K. A. Hambarde, H. Proença, Information retrieval: Recent advances and beyond, IEEE Access 11 (2023) 76581–76604. doi:10.1109/ACCESS.2023.3295776.

[2] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER corpus: English–French parallel data for multilingual wordplay recognition, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.

[3] L. Ermakova, et al., Overview of CLEF 2026 JOKER track: Humor detection, search, and translation, in: M. Hagen, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer-Verlag, 2026.

[4] L. Ermakova, A.-G. Bosser, T. Miller, A. Jatowt, Overview of the CLEF 2024 JOKER task 1: Humour-aware information retrieval, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1775–1785.

[5] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the CLEF 2025 JOKER task 1: Humour-aware information retrieval, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2744–2760.

[6] P. Vachharajani, et al., CLEF 2026 JOKER task 1: Humor-aware information retrieval in English and Hinglish, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[7] Y. Xie, J. Li, P. Pu, Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, 2021, pp. 33–39. URL: https://aclanthology.org/2021.acl-short.6.

[8] E. Schuurman, M. Cazemier, L. Buijs, J. Kamps, University of Amsterdam at the CLEF 2024 JOKER track, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1909–1922.

[9] A. Gepalova, A. Chifu, S. Fournier, CLEF 2024 JOKER task 1: Exploring pun detection using the T5 transformer model, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[10] H. Baguian, N. A. Huynh, JOKER track @ CLEF 2024: The jokesters' approaches for retrieving, classifying, and translating wordplay, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1811–1817.

[11] A. Arampatzis, et al., CLEF 2024 JOKER track results, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[12] T. Chaudhari, A. Vora, S. Hotha, S. Sonawane, PICT at CLEF 2025 JOKER track: Humour-aware information retrieval using BERT-enhanced ensemble methods, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2803–2812.

[13] A. S. Doğruož, S. Sitaram, B. E. Bullock, A. J. Toribio, A survey of code-switching: Linguistic and social perspectives for language technologies, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 1654–1666. URL: https://aclanthology.org/2021.acl-long.131.