From Precision to Perception: User Surveys in the Evaluation of Keyword Extraction Algorithms

Jingwen Cai
jingwen.cai@umu.se
Umeå University
Umeå, Sweden

Sara Leckner
sara.leckner@mau.se
Malmö University
Malmö, Sweden

Johanna Björklund
johanna.bjorklund@umu.se
Umeå University
Umeå, Sweden

ABSTRACT

Stricter regulations on personal data are causing a shift towards contextual advertising, where keywords are used to predict the topical congruence between ads and their surrounding media contexts — an alignment shown to enhance advertising effectiveness. Recent advances in AI, particularly large language models, have improved keyword extraction capabilities but also introduced concerns about computational cost. This study conducts a comparative, survey-based evaluation experiment of three prominent keyword extraction approaches, emphasising user-perceived accuracy and efficiency. Based on responses from 552 participants, the embedding-based approach emerges as the preferred method. The findings underscore the importance of human-in-the-loop evaluation in real-world settings.

CCS CONCEPTS

• **Information systems → Crowd sourcing**; **Information extraction**; *Relevance assessment*; *Retrieval efficiency*.

KEYWORDS

Keyword extraction, Human evaluation, Data analysis, Statistical methods, Word embeddings, Language models

ACM Reference Format:
Jingwen Cai, Sara Leckner, and Johanna Björklund. 2025. From Precision to Perception: User Surveys in the Evaluation of Keyword Extraction Algorithms. In *Workshop on Human-In-the-Loop Data Analytics (HILDA' 25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3736733.3736741

1 INTRODUCTION

Keyword extraction is a central task in many natural language processing (NLP) algorithms, used to extract words or phrases that effectively represent a given document. It underpins applications such as information retrieval, document categorisation, and sentiment analysis [9, 25]. Various algorithmic solutions have been proposed [6, 11, 23, 31], from early statistical methods based on word frequencies to recent implementations using large language models (LLMs), each differing in how they assess word relevance and compose keyword sets. In contextual advertising, keyword extraction plays a crucial role, aligning ads with contextually relevant media content, improving ad targeting, click-through rates and content-moderation effectiveness [18, 28, 37]. However, this domain imposes several unique challenges: real-time auctions demand low-latency processing, and keyword accuracy directly affects revenue potential. Thus, there is a growing need for methods that are not only effective but also computationally efficient.

Despite notable advances, existing evaluation methods tend to rely on static gold-standard datasets created for general purposes, often neglecting user perception and context-specific relevance [3, 35]. This creates a mismatch between algorithmic outputs and practical utility — particularly when semantic alignment is prioritised over surface-level matches. Moreover, while LLM-based methods offer potential gains, their perceived value over simpler alternatives from a user perspective remains under-explored. Therefore, there is a need for increased human involvement in the evaluation and for efficient methods grounded in human-perceived coherence.

To address these gaps, this study uses a human-in-the-loop evaluation framework for comparing keyword extraction approaches. We conduct both quantitative and qualitative assessments of three widely used keyword extraction algorithms, each with a different level of complexity: (i) TF-IDF, a straightforward statistical method; (ii) KeyBERT, which uses a deep neural network for keyword selection; and (iii) Llama 2, a robust open source LLM of medium size. For the quantitative assessment, keywords are benchmarked against participant-selected gold standards, while for the qualitative analysis, user preferences are collected through survey-based user experiments. Novel analytical tools are introduced to control for experimental variability in subjective evaluations.

By emphasising human-in-the-loop evaluation and user-perceived relevance, this study aims to inform both researchers and practitioners in selecting keyword extraction methods suited to real-world applications. Specifically, we address the following research questions:

RQ1 What factors should be considered when conducting human-in-the-loop evaluations of keyword-extraction algorithms?

RQ2 How do the keywords generated by TF-IDF, KeyBERT, and Llama 2 perform in terms of precision-based and perception-based evaluations?

RQ3 What are the implications of the findings for real-world applications such as contextual advertising?

2 THEORETICAL BACKGROUND

This study evaluates three keyword extraction methods: a classical statistical method, a modern extension leveraging neural word embeddings, and a large language model. One of the oldest, but still popular, methods for keyword extraction is TF-IDF [1, 21, 32]. It applies to document collections and estimates the importance of a

HILDA' 25, June 22–27, 2025, Berlin, Germany

ACM ISBN 979-8-4007-1959-2/2025/06
https://doi.org/10.1145/3736733.3736741

word in a given document by relating the frequency of the word within the document (*term frequency, TF*) to the number of documents in the collection that contain the word (*document frequency, DF*). Although TF-IDF is easy to implement and computationally efficient, it fails to distinguish between the alternative meanings of a term. This limits the accuracy of the method. KeyBERT [17], in contrast, uses the pre-trained language model BERT [11] to capture word semantics by translating them into *embeddings* for keyword selections. Prior studies have shown that KeyBERT achieves higher similarity to author-assigned keywords [20]. It also performs better on domain-specific datasets [29], largely due to its ability to reflect the context and semantics of the text, which statistical methods such as TF-IDF ignore [20]. While more accurate, KeyBERT requires more computational resources than TF-IDF and its performance depends on domain and fine-tuning [2, 29, 30]. The third extraction approach is an LLM, here represented by Llama 2 [22]. LLMs have become general-purpose tools in NLP due to their power and flexibility, but at the expense of high processing costs. Llama 2 is an open-weight model trained on public data, and has been fine-tuned through reinforcement learning with human feedback.

Evaluations of keyword extraction algorithms typically rely on automated methods, focusing on precision and recall with respect to gold-standard keywords [3]. This makes comparison possible across studies using the same dataset, but does not capture how well algorithms meet human expectations and needs. As end users of these algorithms or downstream NLP tasks, humans are generally better at understanding fundamental aspects that automated approaches struggle with, such as semantics and context [38], although LLMs have narrowed this gap.

While existing studies focus on incorporating human preferences into model training, for example, through instruction tuning and human-in-the-loop approaches [10, 16, 33], few explicitly analyse user perception of algorithm performance. Comprehensive user evaluation as a means of performance assessment remains under-researched. While some studies have assessed the effectiveness of information retrieval systems [12, 36], and document clustering approaches [13], in terms of human-perceived coherence, such studies are scarce. In the context of keyword extraction, human-in-the-loop evaluations are particularly limited, despite their potential to provide more nuanced and practically relevant insights. This oversight may stem from the practical challenges of manual evaluation processes, which are typically resource-intensive, time-consuming, domain-specific, and not easily reusable [5, 8, 14, 27]. For this reason, there is also a lack of efficient methods to reliably assess human interpretation [13]. Given the importance of keyword extraction in applications like contextual advertising — where user engagement and relevance are key — there is a clear need to assess these methods based on user experience. Human-in-the-loop approaches are essential not only for improving extraction algorithms, but also for developing and refining evaluation methods themselves.

3 METHODOLOGY

To answer the research questions outlined in the introduction, we complement a quantitative benchmark of algorithmic performance with a qualitative evaluation, based on three survey-based user experiments[1]. To address RQ1 and prepare for the experiments, we formulate a set of evaluation properties which are presented in Section 3.1. For RQ2, we quantify the standard quality metrics across the three evaluated algorithms and analyse the experimental results based on the defined properties. Finally, we analyse the outcome of the experiments to answer RQ3.

3.1 Target Properties

Inspired by the works of Stiennon et al. [33] and Shin et al. [30], we identify a set of properties that characterise high-quality keyword sets. To this end, let d be a news article and $W = \{w_1, w_2, \ldots\}$ an ordered set of keywords extracted from d. We say that the set W is a *proper* keyword set for the news article d if it is:

- ***comprehensive*** in that it covers multiple aspects of d,
- ***representative*** in that it captures the central features of d,
- ***distinctive*** in that its keywords stand out as unique for d compared to the other words in d, and
- ***reasonable*** in that it appears coherent and meaningful.

As we will further discuss in Section 3.3, these properties are partially overlapping, but combined they provide a fair predictor of overall quality. To validate this claim, participants were asked to give explicit feedback on the subjective ***overall quality*** of the presented keyword sets. Moreover, during the user evaluation, gold-standard keyword sets were collected and used to quantitatively calculate standard quality metrics in terms of ***precision***, ***recall***, ***cosine similarity*** and ***edit distance***. Ideally, a good algorithm should achieve high precision, recall and cosine similarity, in combination with a small edit distance.

3.2 Experimental Setup

Online news articles were chosen for the user evaluation experiments because they are a common target for contextual advertising and their literary style lends itself well to algorithmic analysis. Articles were sourced from a dataset provided by Aeterna Labs[2], containing over 45 000 news articles from quality publications such as *The Guardian*, mostly published in 2022. After filtering for length (under 350 words), 50 articles were randomly sampled. Finally, five task articles (see Table A.1 in Appendix A) were manually selected to represent a diverse range of topics — society, sports, culture, music, and wildlife — while ensuring the use of accessible language without domain-specific terms. For each task article, ten keywords were generated using TF-IDF, KeyBERT, Llama 2 and participant-annotated "gold standards" (see Table A.2 in Appendix A), creating a 5 × 4 task instance pool.

Data preprocessing used the *ROUGE* list [15] to remove stop words, but was otherwise kept to a minimum to avoid influencing the keyword extraction algorithm comparisons. Keywords were extracted using the default 1-gram settings for all algorithms to adhere to the minimal-variable and standard implementation approach. For KeyBERT, we used the pre-trained model *all-MiniLM-L6-v2* [17] and for Llama 2 we chose the 7 billion parameter version with prompts based on the KeyLLM project[3].

[1] Data and detailed results are available at https://github.com/JingWen17/Precision_to_Perception

[2] Aeterna Labs is a Swedish adtech company; see https://aeternalabs.ai/

[3] See https://maartengr.github.io/KeyBERT/guides/keyllm.html

Participants were recruited from the *Prolific* platform[4]. A total of 552 participants residing in the UK took part in the study across three experiments. Only the input from those who completed the entire task was considered a valid feedback data point and included in the analysis. To maintain anonymity, each participant was assigned a unique identification number and no identifying information was collected. The online survey was designed using the web-based annotation tool *Potato* [26], including four sections: introduction, demographics, main tasks and experience feedback. In the main task section, each participant was randomly assigned several task instances from the pool. Each task instance consisted of reading the task article, reviewing the extracted keywords, and rating them on a seven-point scale using Likert-type questions (see Section 3.3.1).

3.3 Experiments

Three user evaluation experiments were conducted, each following the procedure outlined in Section 3.2. The *preliminary experiment* aimed to test the interactive workflow, refine the phrasing of task questions, and verify technical functionality. The *main experiment* gathered evaluation data on the performance of each extraction method in isolation. To support further analysis of user perceptions, the *supplementary experiment* was then performed to allow a direct comparison between the methods.

3.3.1 *Preliminary Experiment.*

To ensure clarity and reduce the risks of poorly worded questions affecting participants' responses, each of the five properties described in Section 3.1 was assessed using three differently formulated questions:

(1) ***Comprehensiveness***
   (a) How well do the keywords cover the main article?
   (b) To what extent do you think these keywords cover important information from the article?
   (c) To what degree do you think these keywords comprehensively summarise the main article?

(2) ***Representativeness***
   (a) To what extent do the keywords adequately capture the properties of the main article?
   (b) In your opinion, how accurately do these keywords represent the overall article?
   (c) To what degree do you think these keywords represent the original article?

(3) ***Distinctiveness***
   (a) How unique do you find these keywords in relation to the original article?
   (b) How do these keywords stand out when compared to other words in the article?
   (c) How would you rate these keywords in terms of their distinctiveness for the article?

(4) ***Reasonableness***
   (a) In your opinion, does the selection of these keywords seem reasonable in the context of the article?
   (b) If you were asked to select a set of keywords from the article, then how would you rate the reasonableness of the given ones compared to your selection?
   (c) To what degree do you think these keywords make sense in relation to the original article?

(5) ***Overall quality***
   (a) Considering the previous questions, what is your general opinion of these keywords for this article?
   (b) To what extent do you agree with the selection of these keywords based on their overall quality to reproduce the original article?
   (c) How would you evaluate the overall quality of these automatically generated keywords?

In addition, participants were asked to choose ten keywords from each article according to their personal preferences. At the end of the preliminary experiment, gold-standard keyword sets were formed based on the ten most frequently chosen words for each task article. This experiment also asked about the participants' fatigue levels which helped to calibrate the optimal number of task instances assigned to each participant to balance their engagement with annotation efficiency.

The preliminary experiment involved 196 valid participants, each randomly assigned three task instances covering different articles, resulting in approximately 40 data points for each article-algorithm combination. To reduce the mental load, the questions were split into two sets: *(comprehensiveness, distinctiveness, overall quality)* and *(representativeness, reasonableness, overall quality)*. Each task instance comprised only one question set, requiring participants to complete nine Likert ratings and select ten keywords.

To understand the impact of the alternative question phrasings, we developed two novel statistical approaches. In *horizontal analysis*, each combination of task instance (article and algorithm) and property c was assessed individually. For each combination, the consistency of the participants' answers was measured across three different phrasings related to property c. For example, the middle three columns in Figure 1 show the average scores and standard deviations for the three phrasings on distinctiveness. In this case, the participants' scores appear to be fairly independent of the choice of phrasing. It makes little difference whether question 3.a, 3.b, or 3.c is used $\chi^2(df = 38, N = 60) = 3.32$, $p > .99$. However, the phrasing 3.a yields slightly lower standard deviations, indicating more consistent responses.

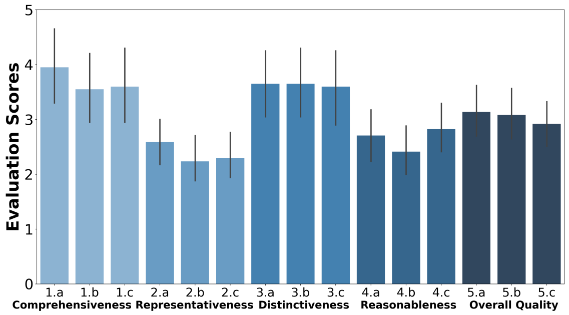

Figure 1: An example of horizontal analysis on feedback consistency for a task instance across 20 participants. The y-axis shows the mean values of the scores awarded by the participants, with the black lines indicating standard deviations.

[4] See https://www.prolific.com

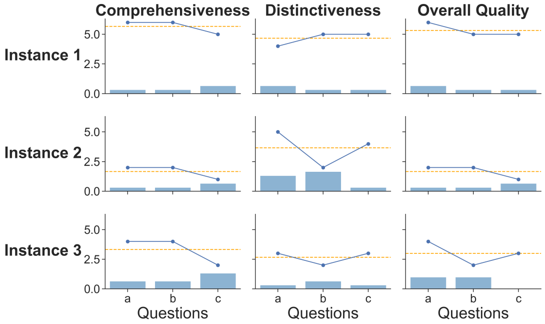

Figure 2: An example of vertical analysis on feedback consistency for a participant. Each column of diagrams corresponds to a property, while each row corresponds to a task instance. The x-axis represents the alternative phrasings for each property. The participant's scores are shown on the y-axis. The horizontal orange lines show the mean scores across three phrasings for the same property and the same instance. The bars show the differences between each score and their corresponding mean score.

In *vertical analysis*, we measured the variation in individual participants' responses. When an individual assesses the same keyword set under the same property but with different phrasings, the average of these scores is assumed to reflect their "true" opinion. Therefore, for all task instances, the smaller the difference between the specific phrasing score and this average, the more indicative that particular phrasing is. Figure 2 shows an example of a participant's ratings and absolute deviations of one set of questions. In this case, for comprehensiveness, phrasing 1.c shows the greatest deviation from the average scores (indicated by the orange lines) across three instances, suggesting it is less reliable and should be avoided in the main study compared to the other two phrasings.

3.3.2 Main Experiment. Based on the horizontal and vertical analyses of the preliminary experiment's results, phrasings 1.a, 2.c, 3.a, 4.a, and 5.c were chosen for each property. Consequently, only these five were used as task questions in the main experiment to make the task more manageable for the participants. In addition to the automatically extracted keyword sets, the gold-standard keyword sets, stemming from participant annotations in the preliminary experiment, were also included without disclosing the fact that they were not algorithmically generated. In this case, each participant was randomly assigned four task instances, one more than in the preliminary experiment, due to the reduced number of task questions and participants' feedback on fatigue levels reported in the preliminary experiment (three instances with ten questions for each are manageable on average). The number of evaluation data points required for each of the 20 task instances in the main experiment was increased to 50 compared to the preliminary experiment, where the focus was on preparation. Thus, a total of 264 participants were included in the main experiment.

3.3.3 Supplementary Experiment. In both the preliminary and main experiments, each task instance involved assessing a single set of keywords in relation to its source article. This design aimed to minimise the independent variables that could influence participants' responses. However, to allow a direct comparison between these keyword extraction methods, a supplementary experiment was conducted in which participants ranked all four keyword sets extracted from the same article, that is, those generated by TF-IDF, KeyBERT, Llama 2, and the gold standard. To avoid potential bias, these keyword sets were randomised and presented as *Group A, Group B, Group C,* and *Group D*, preventing participants from perceiving a particular order among the keyword sets that might influence their evaluations. Instead of using Likert-based questions, participants choose the best and worst sets for each of the first four properties in Section 3.1, using similar phrasings to those in the main experiment. In addition, participants were asked to rank the keyword sets with respect to their *overall qualities*. Each task instance thus includes eight single-choice questions and one sequencing question. A total of 92 participants were included in this experiment, ensuring at least 50 evaluation data points per task article. Each participant completed three task instances, to reduce the potential fatigue caused by the expanded set of questions.

4 RESULTS AND DISCUSSION

This section presents the results of the experiments described in Section 3.3. These are then analysed and the emergent patterns are discussed. We begin by evaluating the soundness of the approach.

4.1 User Evaluations as an Indicator of Algorithm Performance

Following recommendations from prior research [4], we propose combined manual and automated evaluation. To reduce cost and time, we engage a large number of non-experts in our studies, which also reflects real-world end-user scenarios. The challenge of subjectivity is addressed through the inclusion of a user survey. To analyse the effects of different question phrasings in relation to each evaluation property, we use the *horizontal* and *vertical* analysis methods introduced in Section 3.3.1. Both assume that, whether for the same task instance or the same participant, the standard deviation and mean of responses to differently phrased questions assessing the same evaluation property should not vary significantly. Otherwise, the phrasing results in poor consistency in user feedback and should be avoided in the main experiment.

While Kendall's W [19] is a commonly used non-parametric statistic to test assessment consistency and inter-rater reliability often for ordinal ratings, its general use requires that each task instance must be evaluated the same number of times. This condition is not met in our experiments, where we only ensure that each task instance receives more than the baseline number of valid data points defined for the final evaluation; the remainder of the sampling procedure is entirely random. For this reason, we do not apply Kendall's W in our analysis.

4.2 Target Properties Used for Evaluation

RQ1 aims to evaluate keyword sets regarding the properties, tailored to capture fundamental aspects of how humans assess keyword quality. Overall, the experimental results indicate that the properties fulfil their intended purpose. The exception is *distinctiveness*,

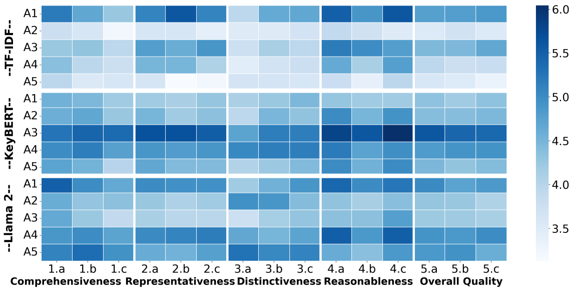

Figure 3: Averaged evaluation scores from the preliminary experiment. The colour of each cell reflects the average score of the corresponding question for participants who were randomly assigned to that task instance. The deeper the colour, the higher the evaluation score.

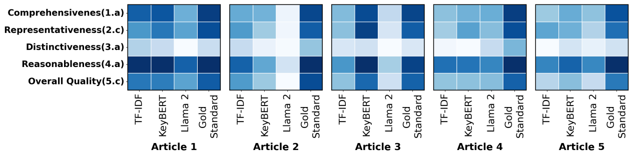

Figure 4: Averaged evaluation scores for each task article from the main experiment. The colour of each cell reflects the average score of the corresponding question across the participants randomly assigned to that task instance. The deeper the colour, the higher the evaluation score.

which receives notably lower participant ratings in both preliminary and main experiments (see Figures 3 and 4). Open-ended feedback reveals that while the keywords effectively cover the general information of the articles, they often miss details such as named entities and the nuanced article's tone. These comments also include gold-standard keywords. For instance, the notably higher *distinctiveness of the gold-standard keywords in Article 4 (see Figure 4) may be attributed to its selection of all specific brand and celebrity names (see Table A.2 in Appendix A), which makes it less general compared to the algorithmic keyword sets.* Another explanation is that, despite the efforts to formulate clear task questions, the concept of distinctiveness may not have been adequately defined, potentially introducing randomness into participant ratings. Addressing the measurement of distinctiveness with greater precision is an intriguing topic for future studies.

The evaluation scores from the main experiment (Figures 4) exhibit clear patterns in participants' assessments of the properties, which repeat across different articles and keyword sets. Specifically, *distinctiveness* consistently receives the lowest scores and *comprehensiveness* and *representativeness* are rated similarly. In addition, there is no explicit evidence of trade-offs among the properties based on evaluation ratings or comments. For instance, in Figure 4, while participants rate Article 1 higher on *distinctiveness* compared to Article 5 on average, the ratings on other properties also increase, suggesting a general trend rather than distinct trade-offs. Further analysis using the Pearson correlation coefficients of the main experiment's results reinforces these findings. The analysis shows a stronger positive correlation among *comprehensiveness*, *representativeness* and *resonableness* than the correlation between the *distinctiveness* with other properties across all keyword sets. For example, *comprehensiveness* and *representativeness* are found to be moderately positively correlated in gold standard ratings, $r(N = 255) = .77, p < .01$, whereas the correlation between *comprehensiveness* and *distinctiveness* are more neutral, $r(N = 255) = .38, p < .01$. Across different task articles, these correlations remain positive but generally decrease. For example, the correlation between *comprehensiveness* and *distinctiveness* drops to $r(N = 208) = .19, p < .01$ in ratings of Article 3.

4.3 The Influence of Algorithmic Complexity on User Preferences

Turning to RQ2, participants' preferences vary notably depending on the task articles. For example in Figure 3, KeyBERT keywords (row 6) are rated lower than TF-IDF (row 1) and Llama 2 (row 11) for Article 1. However, this pattern is completely reversed for Article 3 (rows 3, 8, and 13). Despite this article-specific variation, a slight overall trend emerges in favour of KeyBERT across all articles. On average, KeyBERT receives 14.8% higher ratings than TF-IDF and 1.3% higher than Llama 2. This aligns with other comparative evaluations [2, 20], highlighting KeyBERT's ability to produce high-quality keywords. However, the finding that it performs at least as well as Llama 2 in terms of direct user ratings is novel. Figure 4 further supports this result. For instance, KeyBERT's high ratings for Article 3 are comparable to the gold standard, receiving a 1.2% higher average score. When averaging scores across all articles, KeyBERT keyword sets are rated highest, on average 5.8% higher than the other two algorithms, followed by TF-IDF and then Llama 2. These results suggest that a relatively simple but specialised solution can outperform a more powerful one in terms of user preferences when extracting keywords, especially for structured, high-quality corpora.

To quantitatively assess the quality of the algorithmic keyword sets, they are compared against the gold-standard keywords using standard quality metrics, see Table 1. The first two columns show the *precision* and *recall*, which results in identical figures in this case because both algorithmic and gold-standard sets have a fixed length of ten. Higher precision and recall indicate stronger overlap with human-annotated keywords, reflecting the effectiveness and comprehensiveness of the evaluated algorithm in extracting keywords that closely resemble human choices. To assess the semantic similarities, Word2Vec [24] is used to compute the *cosine distance* between each algorithmic keyword set and the gold-standard keyword set. A

Table 1: The comparisons of algorithmic keywords with gold-standard keywords. The metric values for each keyword extraction method are averaged across five task articles.

Keywords	Precision	Recall	Cosine Similarity	Edit Distance
TF-IDF	0.56	0.56	0.50	3.10
KeyBERT	0.54	0.54	0.55	2.92
Llama 2	0.40	0.40	0.44	2.90

higher cosine distance reflects greater semantic alignments. Finally, the *edit distance* concerning position replacements is calculated, to account for the order of the keywords. Algorithmic keyword orders are based on their weights during extraction, whereas the order of gold-standard keywords reflects the frequency of participant selections. Therefore, even if an algorithm extracted the correct terms, a low ranking of highly participant-selected terms would still increase the edit distance.

Overall, the quantitative analysis combining precision, recall, cosine distance, and edit distance aligns with findings from the preliminary and main experiments. When evaluated against the gold-standard keywords, KeyBERT's keywords are found to be "closer" to those selected by participants, which may explain why they had higher scores. KeyBERT's ability to capture semantics and context has been highlighted in previous work [2, 20]. With respect to the research questions, this makes it more likely that KeyBERT will resonate with a target audience and align with the contextual nuances of advertising campaigns, which in turn increases the likeliness for advertising efficiency and hence the return on investment [7, 34].

4.4 Automatically Versus Manually Selected Keywords: The Gap Is Still Significant

In the main experiment, participants occasionally assign higher scores to an algorithmic keyword set than gold-standard keywords. However, in the supplementary experiment, where participants rank all four keyword sets directly, the gold-standard keywords almost always come out on top (see Figure 5). There is a marked preference for gold-standard keywords, where the total number of participants selecting them as the best set is typically 1.5 to 4.6 times higher than algorithmic keyword sets. However, for members of the latter set, the differences are not statistically significant, $\chi^2(df = 14, N = 1493) = 8.53, p = .86$. Related to RQ1 and RQ3, this is further evidence of the importance of human annotations in algorithm evaluations. For contextual advertising, this also shows the importance of including manually entered keywords in targeting, and taking a human-in-the-loop approach to AI-driven advertising.

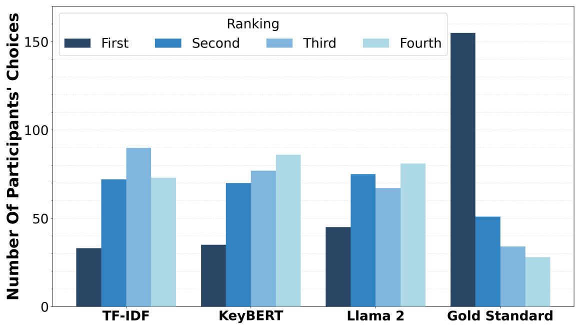

Figure 5: Participant rankings of the keyword sets. The x-axis represents each extraction method, and the y-axis shows the frequency with which each keyword set is chosen for a specific rank.

5 CONCLUSION

Keyword extraction remains a vital task in NLP. Recent advances in language modelling contribute to groundbreaking progress in a wide range of areas, but come at the cost of increased computational complexity and higher hardware requirements. Although neural models are often fine-tuned with human feedback, less effort has been spent on analysing the user-perceived performance of downstream algorithms that depend on the algorithms or LLM's output. For keyword extraction, simpler models can still be competitive, especially for applications such as contextual advertising that require the efficient processing of large amounts of text.

Beyond typical precision-based evaluation metrics, this study emphasises user perceptions in assessing three representative keyword extraction methods: TF-IDF, KeyBERT, and Llama 2. To ensure consistent feedback and minimise unexpected shifts in user responses, a set of evaluation properties is developed and implemented. In addition, two novel analytical approaches — horizontal and vertical analysis — are introduced for analysing question phrasing consistency, offering valuable insights for improving experimental design. The experimental results reveal variations in the participants' preferences across different articles. Overall, KeyBERT outperforms TF-IDF and Llama 2, proving its effectiveness. Although the gold standard is strongly preferred by participants, differences among the algorithmic sets are not statistically significant. These findings underscore the importance of human-in-the-loop evaluation and provide insights for researchers and practitioners to reflect on the balance between performance optimisation and user-centred evaluation, particularly for real-world applications where computational efficiency is critical and the primary goal is to meet end-user needs. Rather than solely pursuing incremental improvements in precision, it is worth considering what real users truly value in practice.

6 LIMITATION AND CHALLENGES

This study focuses on comparing different algorithm-generated keywords by surveying participants on predefined evaluation properties. Future work could further investigate participants' perceptions of the evaluation properties. To minimise task fatigue and focus on our initial research direction, this study is limited to three mainstream algorithms. Nonetheless, they represent key techniques and stimulate further research attention towards practical user perspectives in this area. Future studies can expand this work by exploring additional approaches, such as unsupervised graph-based algorithms like TextRank and other LLMs like Llama3 and Mistral.

Keywords play a crucial role in contextual advertising targeting. Although this study is motivated by such practical applications, its broader aim is to highlight the importance of incorporating user perceptions in algorithm evaluations. It would be interesting to compare algorithm-generated keywords with user-selected keywords in real-world recommendation scenarios, such as contextual advertising, and to investigate how the quality of the extracted keywords affects users' perceived relevance between articles and ads, and more practically, their purchasing intentions. However, there are challenges in developing such experiments. Factors such as numerous external variables like ad layouts and difficulties in accurately capturing users' true purchasing intentions via surveys are likely to influence the experimental outcomes.

REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing & Management* 39, 1 (2003), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3 https://doi.org/10.1016/S0306-4573(02)00021-3.

[2] Zaira Hassan Amur, Yew Kwang Hooi, and Gul Muhammad Soomro. 2022. Automatic Short Answer Grading (ASAG) using Attention-Based Deep Learning MODEL. In *Proceedings of the 2022 International Conference on Digital Transformation and Intelligence (ICDI)*. IEEE, Sarawak, Malaysia, 1–7. https://doi.org/10.1109/ICDI57181.2022.10007187.

[3] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36 (2015), 15–24. https://api.semanticscholar.org/CorpusID:6134326 https://doi.org/10.1609/aimag.v36i1.2564.

[4] Gábor Berend and Veronika Vincze. 2012. How to evaluate opinionated keyphrase extraction?. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics, Jeju, Republic of Korea, 99–103. https://aclanthology.org/W12-3715.

[5] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4. Association for Computing Machinery, New York, USA, Article 93, 26 pages. Issue CSCW2. https://doi.org/10.1145/3415164 https://doi.org/10.1145/3415164.

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509, C (2020), 257–289. https://doi.org/10.1016/j.ins.2019.09.013 https://doi.org/10.1016/j.ins.2019.09.013.

[7] Jinyuan Chen, Haitao Zheng, Yong Jiang, Shutao Xia, and Congzhi Zhao. 2019. A probabilistic model for semantic advertising. *Knowledge and Information Systems* 59, 2 (2019), 387–412. https://doi.org/10.1007/s10115-018-1160-7 https://doi.org/10.1007/s10115-018-1160-7.

[8] Chenghan Chiang and Hungyi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics, Toronto, Canada, 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870 https://aclanthology.org/2023.acl-long.870.

[9] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: Evolution of research methods and topics. *Artificial Intelligence Review* 56, 8 (2023), 8469–8510. https://doi.org/10.1007/s10462-022-10386-z.

[10] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* 15 (2024), 1–14. https://doi.org/10.1038/s41467-024-45563-x https://doi.org/10.1038/s41467-024-45563-x.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423 https://doi.org/10.18653/v1/N19-1423.

[12] Alberto Diaz, Antonio Garcia, and Pablo Gervas. 2007. User-centred versus system-centred evaluation of a personalization systems. *Information Processing and Management* 44, 3 (2007), 1293–1307. https://doi.org/10.1016/j.ipm.2007.08.001.

[13] Anton Eklund, Mona Forsmans, and Frank Drewes. 2025. Comparing Human-Perceived Cluster Characteristics through the Lens of CHIPE: Measuring Coherence beoynd Keywords. *Journal of Data Mining and Digital Humanities* NLP4DH (2025). https://doi.org/10.46298/jdmdh.15044.

[14] Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information Processing & Management* 56, 5 (2019), 1794–1814. https://doi.org/10.1016/j.ipm.2019.04.001 https://doi.org/10.1016/j.ipm.2019.04.001.

[15] Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint* (2018). https://doi.org/10.48550/arXiv.1803.01937.

[16] Amelia Glaese, Nathan McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, A. See, Sumanth Dathathri, Rory Greig, Charlie Chen, ..., and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint* (2022). https://doi.org/10.48550/arXiv.2209.14375.

[17] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. v0.3.0 [software], Zenodo. https://doi.org/10.5281/zenodo.4461265 https://doi.org/10.5281/zenodo.4461265.

[18] Emil Häglund and Johanna Björklund. 2024. AI-Driven Contextual Advertising: Towards relevant messaging without personal data. *Journal of Current Issues and Research in Advertising* 45, 3 (2024), 301–319. https://doi.org/10.1080/10641734.2024.2334939.

[19] Maurice G. Kendall. 1949. *Rank Correlation Methods* (first ed.). Griffin.

[20] Muhammad Q. Khan, Abdul Shahid, M. Irfan Uddin, Muhammad Roman, Abdullah Alharbi, Wael Alosaimi, Jameel Almalki, and Saeed M. Alshahrani. 2022. Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science* 8:e967 (2022). https://doi.org/10.7717/peerj-cs.967 https://doi.org/10.7717/peerj-cs.967.

[21] Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development* 1, 4 (1957), 309–317. https://doi.org/10.1147/rd.14.0309 https://doi.org/10.1147/rd.14.0309.

[22] Meta. 2023. Introducing LLaMA: A foundational, 65-billion-parameter large language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Online.

[23] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252 https://aclanthology.org/W04-3252.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint* (2013). https://doi.org/10.48550/arXiv.1301.3781.

[25] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2019. Textual keyword extraction and summarization: State-of-the-art. *Information Processing & Management* 56, 6 (2019), 102088. https://doi.org/10.1016/j.ipm.2019.102088 https://doi.org/10.1016/j.ipm.2019.102088.

[26] Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Abu Dhabi, UAE, 327–337. https://doi.org/10.18653/v1/2022.emnlp-demos.33 https://doi.org/10.18653/v1/2022.emnlp-demos.33.

[27] Kevin Roitero, Andrea Brunello, Giuseppe Serra, and Stefano Mizzarov. 2020. Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations. *Information Processing & Managament* 57, 2 (2020), 102149. https://doi.org/10.1016/j.ipm.2019.102149 https://doi.org/10.1016/j.ipm.2019.102149.

[28] Joni Salminen, Bernard J. Jansen, and Mekhail Mustak. 2023. How Feature Changes of a Dominant Ad Platform Shape Advertisers' Human Agency. *International Journal of Electronic Commerce* 27, 1 (2023), 3–35. https://doi.org/10.1080/10864415.2022.2158594 https://doi.org/10.1080/10864415.2022.2158594.

[29] Jill Sammet and Ralf Krestel. 2023. Domain-Specific Keyword Extraction using BERT. In *Proceedings of the 4th Conference on Language, Data and Knowledge*. NOVA CLUNL, Vienna, Austria, 659–665. https://aclanthology.org/2023.ldk-1.72 https://aclanthology.org/2023.ldk-1.72.

[30] Hunsik Shin, Hye Jin Lee, and Sungzoon Cho. 2023. General-use unsupervised keyword extraction model for keyword analysis. *Expert Systems with Applications* 233 (2023), 120889. https://doi.org/10.1016/j.eswa.2023.120889 https://doi.org/10.1016/j.eswa.2023.120889.

[31] Mingyang Song, Yi Feng, and Liping Jing. 2023. A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia, 2153–2164. https://doi.org/10.18653/v1/2023.findings-eacl.161 https://doi.org/10.18653/v1/2023.findings-eacl.161.

[32] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21. https://doi.org/10.1108/eb026526.

[33] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize with human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[34] Yajuan Wang, Zhanghua Zhou, and Chonghuan Xu. 2022. The effect of contextual mobile advertising on purchase intention: The moderating role of extroversion and neuroticism. *Frontiers in Psychology* 13 (2022), 1–12. https://doi.org/10.3389/fpsyg.2022.849369.

[35] Lars Wißler, Mohammed Almashraee, Dagmar Monett, and Adrian Paschke. 2014. The Gold Standard in Corpus Annotation. In *5th IEEE Germany Student Conference*, Vol. 21. IEEE, Passau, Germany. https://doi.org/10.13140/2.1.4316.3523 https://doi.org/10.13140/2.1.4316.3523.

[36] Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. 2023. On the role of human and machine metadata in relevance judgment tasks. *Information Processing & Managament* 60, 2 (2023), 103177. https://doi.org/10.1016/j.ipm.2022.103177 https://doi.org/10.1016/j.ipm.2022.103177.

[37] Yanwu Yang and Huiran Li. 2023. Keyword decisions in sponsored search advertising: A literature review and research agenda. *Information Processing &*

Management 60, 1 (2023), 103142. https://doi.org/10.1016/j.ipm.2022.103142 https://doi.org/10.1016/j.ipm.2022.103142.

[38] Torsten Zesch and Iryna Gurevych. 2009. Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, Borovets, Bulgaria, 484–489. https://aclanthology.org/R09-1086 https://aclanthology.org/R09-1086.

A TASK ARTICLES AND KEYWORD LISTS

The URLs of the five task articles and their keywords generated by TF-IDF, KeyBERT and Llama 2 are listed, along with the gold-standard keywords selected by the participants from the preliminary experiment for each article.

Table A.1: URLs of the five sampled task articles.

Id	URL
1	https://www.theguardian.com/society/2022/may/25/more-than-one-in-10-young-women-now-identify-lesbian-gay-bisexual-or-other
2	https://www.theguardian.com/news/audio/2022/jul/29/euro-2022-and-the-future-of-womens-football
3	https://www.theguardian.com/world/2022/may/31/14th-century-samurai-sword-found-in-car-at-swiss-border
4	https://www.theguardian.com/music/2022/oct/26/kanye-west-escorted-out-skechers
5	https://www.theguardian.com/world/2022/jul/21/an-an-worlds-oldest-captive-male-giant-panda-dies-in-hong-kong-zoo-aged-35

Table A.2: The keywords generated by TF-IDF, KeyBERT, Llama 2 and participants for five sampled articles.

Article	TF-IDF	KeyBERT	Llama 2	Gold Standard
1	people, gay, bisexual, uk, heterosexual, identifying, lesbian, figure, identify, young	heterosexual, bisexual, lesbian, gay, gender, straight, sexual, percentage, ethnicity, footballer	lesbian, gay, bisexual, other, female, young, men, age, sexual, orientation	bisexual, lesbian, gay, uk, women, young, orientation, sexual, openness, heterosexual
2	football, people, uk, girl, game, woman, moore, muslim, organisation, passion	lioness, footballer, football, sport, england, woman, girl, playing, fa, hannah	success, girls, uk, guardian, suzanne, football, fa, trolling, spain, brighton	football, lionesses, women, uk, accessible, muslim, girls, role, model, media
3	custom, swiss, sword, authority, driver, franc, found, object, antique, investigation	katana, sword, custom, swiss, antique, smuggled, franc, fine, samurai, zurich	swiss, authorities, sword, japan, discovered, book, contract, invoice, driver, daughter	swiss, sword, customs, smuggled, fines, investigation, criminal, antique, samurai, authorities
4	ye, skechers, antisemitic, recent, comment, company, adidas, gap, longer, longtime	kanye, adidas, rapper, supremacist, footwear, skechers, fashion, brand, white, gap	skechers, kanye, west, unannounced, offices, artist, fashion, statement, brand, antisemitic	skechers, antisemitic, kanye, west, adidas, ye, gap, supremacists, unauthorized, conspiracy
5	park, ocean, kong, chinese, year, giant, hong, panda, friendship, died	panda, jia, chinese, zoo, kong, hong, sichuan, oldest, ocean, ying	panda, oldest, hong, an, died, thursday, zoo, china, jia, ying	panda, kong, hong, oldest, died, zoo, an, euthanised, ocean, park