International Journal of Electrical, Electronics and Computers
Vol-8, Issue-4 | Jul-Aug, 2023
Available: https://aipublications.com/ijeec/
Peer-Reviewed Journal

Low Resource Domain Subjective Context Feature Extraction via Thematic Meta-learning

Vishesh Agarwal[1], Anil Goplani[2], Mohit Kumar Barai[3], Arindam Sarkar[4], Subhasis Sanyal[5]

[1]SQE, Samsung Research Institute Noida, India
Email: v12.agarwal@samsung.com
[2]SQE, Samsung Research Institute Noida, India
Email: anil.goplani@samsung.com
[3]SQE, Samsung Research Institute Noida, India
Email: m.barai@samsung.com
[4]SQE, Samsung Research Institute Noida, India
Email: arindam.s@samsung.com
[5]SQE, Samsung Research Institute Noida, India
Email: s.sanyal@samsung.com

Received: 28 Jun Sep 2023; Accepted:25 Jul 2023; Date of Publication: 02 Aug 2023

Abstract— *The volume of the data is directly proportional to the model's accuracy in data analytics for any particular domain. Once a developing field or discipline becomes apparent, the scarcity of the data volume becomes a challenging proponent for the correctness of a model and prediction. In the proposed state-of-the-art, a transitive empirical method has been used within the same contextual domain to extract features from a low-resource part via a heterogeneous field with factual data. Even though an example of text processing has been used for brevity, it is not limited. The success rate of the proposed model is 78.37%, considering model performance. But when considering human subject matter experts, the accuracy is 81.2%.*

Keywords— ***Data Analytics, Feature Extraction, Feedback review, Natural Language Processing, Text Processing.***

I. INTRODUCTION

The nature of universal events is Volatile, Uncertain, Complex, and Ambiguous [1]. All of these dimensions, as mentioned above, bring a novel context or topic. Some of which may have a positive impact and some negative. For example, the COVID-19 health crisis across the world has affected many lives and occupations. Nassim Nicholas Taleb, in 2007 proposed the 'Black swan theory. He stated," A black swan is an unpredictable event beyond what is typically expected of a situation and has potentially severe consequences. Black swan events are characterized by their extreme rarity, powerful impact, and the widespread insistence they were apparent in hindsight." The question remains can we predict the characteristics of these events? Can we know the unknown when the event is in a nascent state? The quantity and quality of the data play a significant part. Data collection is an ongoing iterative process by which data is continuously collected and analyzed to draw inductive inferences, driven mainly by subjective interpretation of the probability based on past events/prior knowledge [2,3]. But when a limited amount of target domain data is present for adaptation of a model and learning, the prediction and model become undetermined. Data Augmentation is a technique that enhances the quantity and quality of training datasets so that better Learning models can be built [4,5]. The data argumentation technique in Natural Language Processing (NLP) is novel. Mainly data Augmentation algorithms establish synthetic data from an available dataset., But Data argumentation in the field of NLP is intricate compared to other forms of Data Augmentation. For instance, changing the order of words can completely alter the sentence's meaning. For example,' I had my house built' differs from' I had built my house.

Also, the same word can be utilized as an adjective or a noun. Like,' I was traveling through windy road.' 'Windy' can be interpreted as an adjective or a noun (name of a road). From here, we can say context becomes very important. In our research, we have found out that if we can obtain the context of the low-resource domain, then by using other homogeneous context-driven fields where data is copious, we can perform data augmentation, which can be helpful for feature extraction of that low-resource domain. Identifying the context or topic of the lower resource domain is paramount for our research. Topic modeling is a method to find a group of a word associated with pre-learned topics or context [6,7]. A universal set drives each topic or context. Below is an example of a global feedback domain and other probable sub-sets of classes.

II. LITERATURE REVIEW

Lack of data or labeled data is pertinent for low-resource domain feature extraction. Many methods are postulated. The fundamental objective of these studies was based on distant supervision and transfer learning which reduces the need for target supervision [8]. Degrees of freedom are a salient concept in data analytics while considering knowledge discovery in low-resource domain space. Degrees of Freedom are correlated with the maximum number of logically independent values, which can be referred to as a feature in the context of feature extraction. Mintz et al. proposed a Distant supervision method that extracted low-domain resource features using Named Entity Recognition or Relation Extraction. They have used complex knowledge bases like Wikipedia for relational inference [9]. The challenge while using a massive database like Wikipedia is processing time. Another type of method was provided by many researchers based on setting up some labeling rules on low-resource data. They have used various domain experts to create a statistical rule for gaining a transfer learning insight. Recently, the use of deep neural networks has also been proposed for label rules [10,11,12].

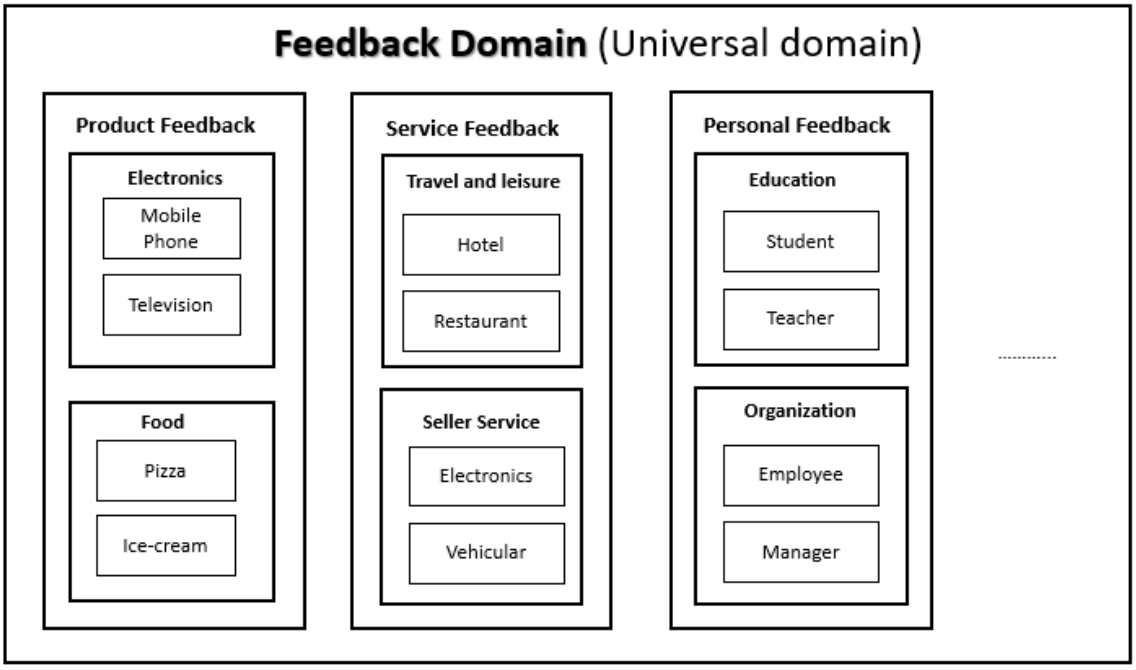


Fig 1. Global Feedback domains

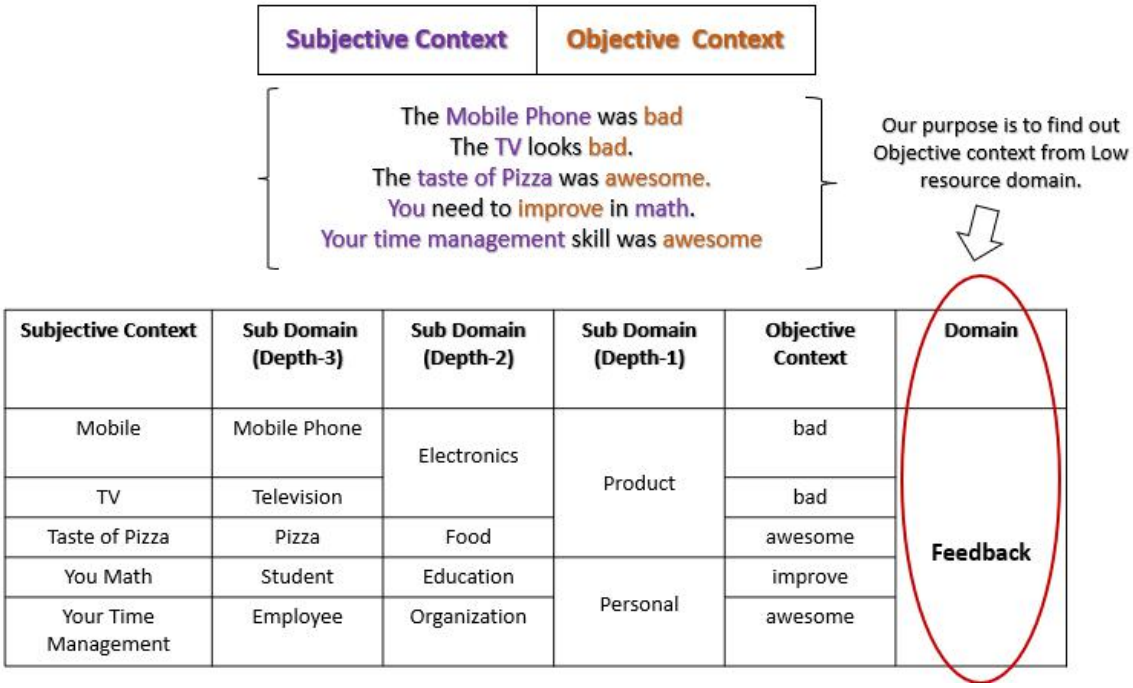
In another work, Cross-Lingual Projections were considered where the task is well supported in one language but not another [13,14,15]. With the advancement of Pre-Trained Transformers via the deep neural network, many researchers have suggested various context-aware word representations that can predict the succeeding word in the sentence. According to them, this can be helpful to obtain features or context from the low-resource domain without substantial task-specific architecture modifications. A deep neural model like BERT or RoBERTa can provide significantly higher accuracy in this context [16,17,18]. Another approach was proposed by Park et al. [19], transferring the knowledge from high-resource domains to low-resource domains using meta-learning. Minimal studies emphasize sharing the knowledge from the high-resource corpora with the low-resource one. Several models [20,21] show better performances than when trained with the low-resource corpora only. But these approaches become conducive in limited scenarios where one or both source and target domains consist of a parallel corpus. In the case of novel subjective domain ushers, these methods fail to predict the domain's probable feature due to the data's

ISSN: 2456-2319
https://dx.doi.org/10.22161/eec.84.1

unavailability. In our proposed method, dividing the text corpora into subjective and objective contexts, we extract knowledge information using cooccurrence statistical relations based on objective context. Then utilize these transitive inference statistics as the input of the embedding model to learn inference rules for low resource domain. The novelty of our work is based on the derivation of the subjective context feature of a low-resource domain based on transferring knowledge between objective context shared by both high and low-resource domains.

III. METHODOLOGY

The subjective-objective dichotomy is associated with human perception and philosophy. Subjective context is cognate with the objective context. Objectivity is associated with something the same for everyone, while subjectivity refers to something different. Both subjective and objective realism is already manifested in humans. So pertaining to this logical reducibility, we can extrapolate any human-generated speech, Text, Image, etc., which explains some forms of communication can be categorized into subjective and objective contexts or topics. Knowledge discovery in an objective context becomes convenient through transfer learning with the same objective domain, irrespective of the subject. Our work is based on the hypothesis mentioned above. Topic modeling is paramount for knowing the objective context association [22,23]. For this reason, we have used the Latent Dirichlet Allocation (LDA) model, one of the most popular in this field. Researchers have proposed various models based on the LDA in topic modeling.



Subjective Context	Sub Domain (Depth-3)	Sub Domain (Depth-2)	Sub Domain (Depth-1)	Objective Context	Domain
Mobile	Mobile Phone	Electronics	Product	bad	Feedback
TV	Television			bad	
Taste of Pizza	Pizza	Food		awesome	
You Math	Student	Education	Personal	improve	
Your Time Management	Employee	Organization		awesome	

Fig 2. Subjective and Objective context illustration

The purpose of this model is to classify text in a document, for our case, low resource unknown and resource knew heavy domain to a particular topic which is nothing but objective context. LDA builds a topic-per-document and words-per-topic model, modeled according to Dirichlet distributions. The Dirichlet distribution is a Beta distribution with multivariate generalization. The primary motivation concerning LDA is that a corpus is a combination of topics, in our case, Objective Context (OCt), and each topic is a combination of Certain words. For Feedback related objective context, we can find a term like good, excellent, evil, etc. Now LDA uses two types of probabilities: First, the likelihood of words in Corpora d currently assigned to topic OCt. Second, the possibility of assignment of topic OCt to overall corpora. Once the homogeneous Objective context has been obtained for low-resource unknown and known domains, we can take this discovery into the next processing phase, where data cleaning is followed by Noun, Adjective, and Verb parts of speech tagging. In one of their research works, Barai et al. [21] proposed a graph mining technique for domain-specific key feature extraction based on the relation between words surrounding an aspect. Transferring this knowledge to our work between low resource domain and data have resource domain connected by the same objective context, we can observe a transitive relation among both subjective domain contexts. For a better understanding, the below figure has been given,

From this transitive relation, we can undoubtedly extract unknown subjective domains feature via Noun or verb entities.

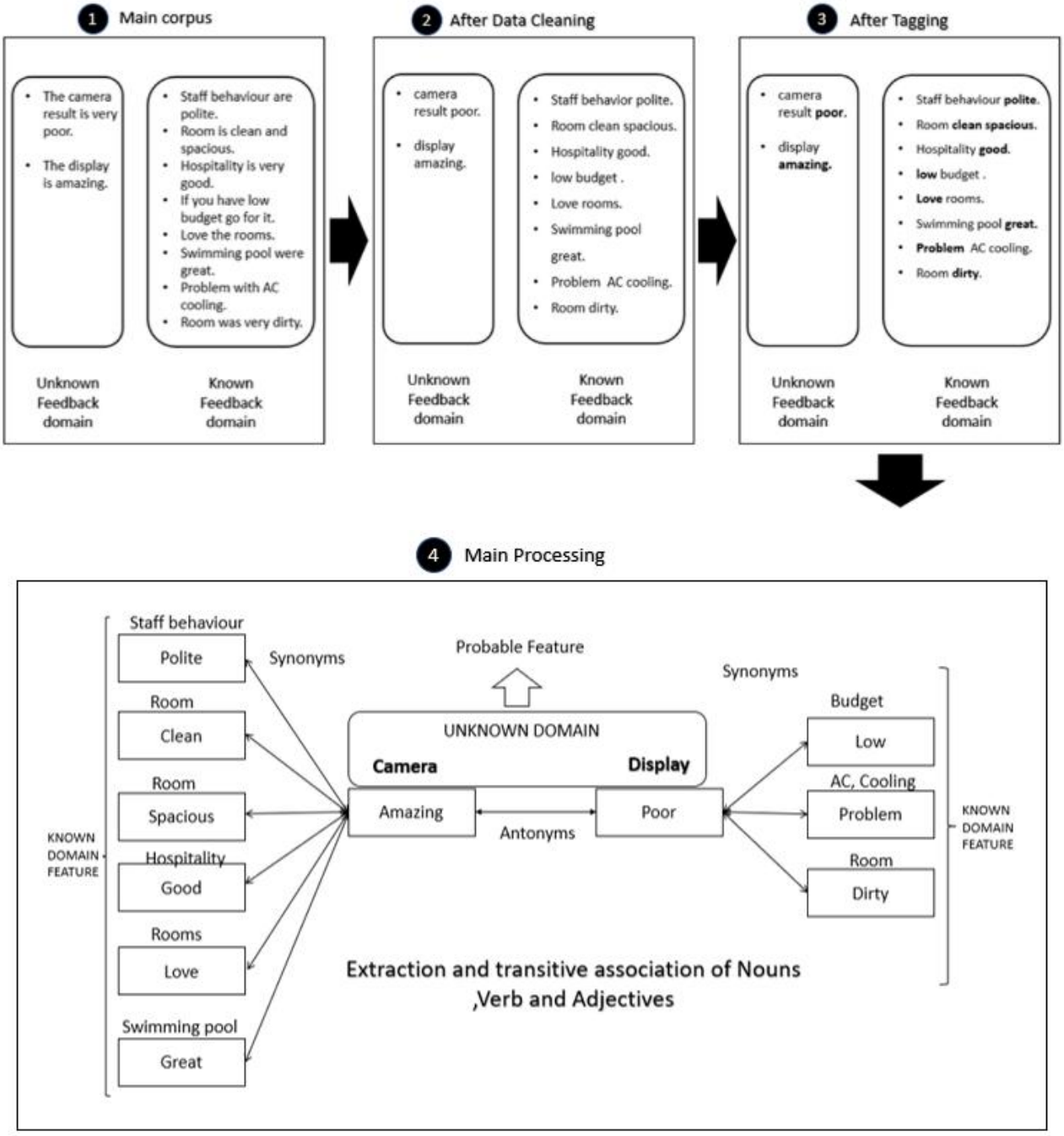


Fig 3. Overall Process illustration

ISSN: 2456-2319
https://dx.doi.org/10.22161/eec.84.1

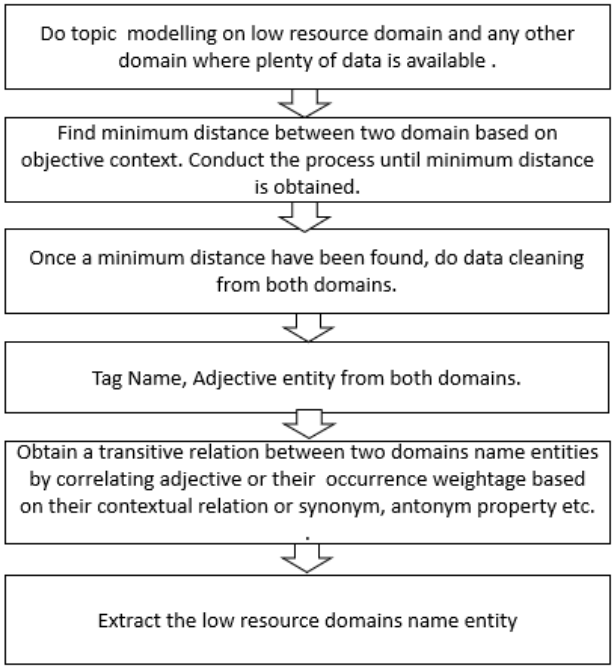


Fig 4. Summarized Algorithm

The mathematical model for our proposal is given below.

Let, $D = \{x | O_x \neq \phi \wedge S_x \neq \phi\}$ Where D is the set of all possible Subjects for which data sets are available in the form of opinion or Feedback.

O_x: set of all objective features of a particular subject "x."

S_x : set of all subjective features of a particular subject "x."

Also, $O_x \cap S_x = \phi$.

The data set F_x for a known subject x will always be a relation and subset of a Cartesian product of S_x and O_x.

$$F_x \subseteq O_x \times S_x$$

Also, $F_x = \{(a,b) | \, dis(a,b) = k\}, Where \; k \in [0, \infty)$

If we have data set available for another subject "y" with unbaled, low resource domain data, we can transitively derive the elements of the subjective set S_y using the known relation F_x .

$$\boldsymbol{F_y = \{(c,d) | \, dis(a,b) = l\}, Where \; l \in [0, \infty)}$$

$$\boldsymbol{S_y = \{c | dis(b,d) \leq \varepsilon \; \forall \; (a,b) \in F_x \wedge \forall (c,d) \in F_y\}}$$

IV. RESULT DISCUSSION

We have kept the Heavy resource objective context domain for our research as Feedback for brevity. After doing the topic modeling based on objective context on both domains, we observed the result below indicated in Fig.5.

Now, we have tried to find the coon features from both objective domains. A total of 32 features were found, containing 87.23% of standard objective features, using obtained objective features in the resource-heavy domain. We have obtained the distance of the named entity in the resource-heavy domain. After that, we optimized the distance based on the occurrence frequency. The same optimized distance has been used in the low-resource domain. And our model accuracy was 78.37%, and once we had validated the data with a subject matter expert, we found out our model accuracy was 80.2%.

V. CONCLUSION AND FUTURE WORK

We have proposed a novel meta-learning model where we have transitively augmented the objective knowledge of a low resource domain field via a data reach homogenous data reach domain to extract probable subjective context features. We can use our method from our research work to extract specific knowledge if a nascent subjective context may be pertinent to lesser unstructured knowledge. In the future, we will try to use our method not only in the case of homogeneous data types (like the text that we did over here) but also in Heterogeneous datatypes.

REFERENCES

[1] Bennett, Nathan & Lemoine, G. James. (2014). What VUCA really means for you. Harvard business review. 92.

[2] Shamoo, Adil & Resnik, David. (2007). Responsible Conduct of Research. Journal of biomedical optics. 12. 39901. 10.1117/1.2749726.

[3] Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3

[4] Shorten, C., Khoshgoftaar, T.M. & Furht, B. Text Data Augmentation for Deep Learning. J Big Data 8, 101 (2021). https://doi.org/10.1186/s40537-021-00492-0

[5] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[6] Miriam, P. . (2012). "Very basic strategies for interpreting results from the topic modeling tool," in Miriam Posner's Blog.

[7] Nugroho, R., Paris, C., Nepal, S. et al. A survey of recent methods on deriving topics from Twitter: algorithm to evaluation. Knowl Inf Syst 62, 2485–2519 (2020). https://doi.org/10.1007/s10115-019-01429-z

[8] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language

Processing in Low-Resource Scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2545–2568, Online. Association for Computational Linguistics.

[9] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

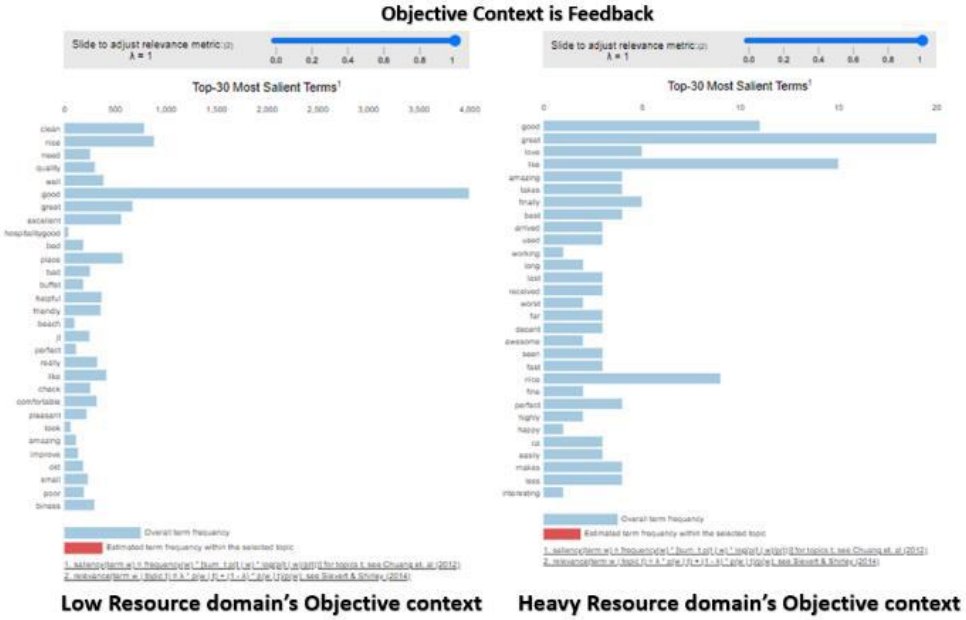


Fig 5. Result

[10] Strötgen, J., Gertz, M. Multilingual and cross-domain temporal tagging. Lang Resources & Evaluation 47, 269–298 (2013). https://doi.org/10.1007/s10579-012-9179-y

[11] Ratner, Alexander & Bach, Stephen & Ehrenberg, Henry & Fries, Jason & Wu, Sen & Ré, Christopher. (2020). Snorkel: rapid training data creation with weak supervision. The VLDB Journal. 29. 10.1007/s00778-019-00552-1.

[12] David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In Proceedings of the First International Conference on Human Language Technology Research.

[13] 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[14] Sostaric, Margita & Pavlović, Nataša & Boltuzic, Filip. (2019). Domain Adaptation for Machine Translation Involving a Low-Resource Language: Google AutoML vs. from-scratch NMT Systems.

[15] Asmussen, C.B., Møller, C. Smart literature review: a practical topic modeling approach to exploratory literature review. J Big Data 6, 93 (2019). https://doi.org/10.1186/s40537-019-0255-7 arXiv:1711.04305 [cs.IR]

[16] Mohit Kumar Barai, Subhasis Sanyal, (2021), DOMAIN SPECIFIC KEY FEATURE EXTRACTION USING KNOWLEDGE GRAPH MINING, Multiple Criteria Decision Making (15), pp. 1-22