2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Construction of domain-specific lexicons based on term statistics

Rafael Rojas-Hernández[*], Asdrúbal López-Chau[*], David Valle-Cruz[†],
Valentin Trujillo-Mora[*], and Elvira Ivone González-Jaimes[*]
[*]*CU UAEM Zumpango*, *Universidad Autónoma del Estado de México*, Zumpango, México
[†]*CU UAEM Tianguistenco*, *Universidad Autónoma del Estado de México*, Tianguistenco, México
Email: {rrojashe, alchau, davacr, vtrujillom, eigonzalezj}@uaemex.mx

***Abstract*—Lexicons are a fundamental resource for sentiment analysis, offensive language identification, trend detection, and document classification. Lexicons have the advantage of being easy to use, but most of the existing lexicons have been created manually. Recently, researchers have been interested in extending the use of lexicons to different fields. In this paper, an easy-to-compute statistics-based method for extracting lexicons in specific domains or ad-hoc lexicons is shown. The proposed method was evaluated on two datasets and achieved 80% accuracy in document classification. This novel approach is expected to be a valuable tool for researchers and practitioners who need to quickly and efficiently create domain-specific and ad-hoc lexicons.**

***Index Terms*—lexicon extraction, custom lexicon, ad-hoc lexicon, specific domain lexicon**

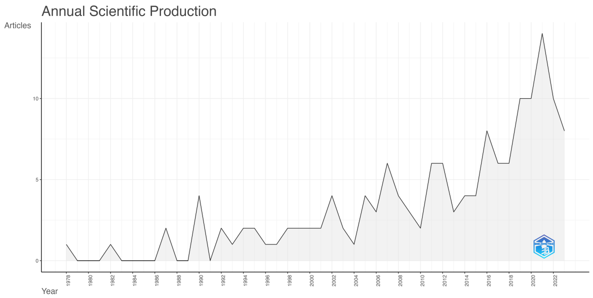

Fig. 1: Annual production of scientific papers on lexicon extraction

I. Introduction

Lexicons are an important resource in sentiment analysis. In simple terms, lexicons are lists of representative terms of a topic. Lexicons can be scored and unscored, being the latter the most frequent. Scored lexicons have a real or integer number associated to each term (which is an indicator of its intensity or importance), unscored lexicons does not have any values associated to the terms.

Lexicons have been applied for sentiment analysis and hate speech identification, among other applications. The ease of using lexicons in these scenarios is that complex model training is not required, compared to machine learning or deep learning based methods for the same purpose. The simplest way to analyze a document with a lexicon is to identify the terms that appear in both, and calculate some basic statistics.

Currently, there are several lexicons focused on sentiment analysis, hate speech or financial topics. The vast majority of these lexicons have been created with the intervention of a group of experts, who analyze documents and extract the most representative terms. This process, in addition to being expensive, is not easily reproducible. Furthermore, lexicons generated in this way cannot be easily expanded, nor they can be successfully applied to other contexts. Other issue is that most of lexicons are for the English language.

In this paper, we propose and evaluate a method to create a domain-specific or ad-hoc scored lexicon from a given corpus. The method is based on basic statistics and set operations.

The contributions of this paper are the following:

- A method to create an ad-hoc lexicon from a corpus.
- The description of the ideal characteristics of the terms in a lexicon.
- The use of nine term features to be used as scores in lexicon.
- Identification of the class imbalance problem in score lexicons, and a proposal of a method to face the problem.

The rest of this paper is organized as follows. Section II shows a literature review on works about lexicon extraction. Section III explain the ideal characteristics of corpora and lexicons. Section IV presents the methodology for creation of ad-hoc or domain specific lexicons. The results obtained from publicly available corpora are shown in Section V. The conclusions are in the last part of this paper.

II. Related works

The field of lexicon construction has evolved significantly since its inception in 1978. Over the years, research efforts have increased remarkably, and the scope of investigations has been expanding. 2021 is a culminating year in the history of lexicon construction research (Figure 1).

Another important aspect of the state of the art in lexicon construction is the wide range of topics investigated. Researchers have worked in a variety of fields, including semantics, sentiment analysis, linguistics, natural language processing, data mining, information retrieval, computational linguistics, learning systems, social networks, and deep learning (Figure 2). This varying focus indicates the interdisciplinary nature of the structure of lexicons, as they contain

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

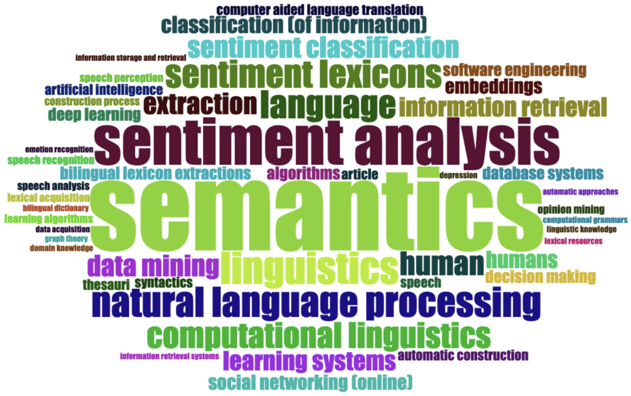

Fig. 2: Topics related to lexicon construction

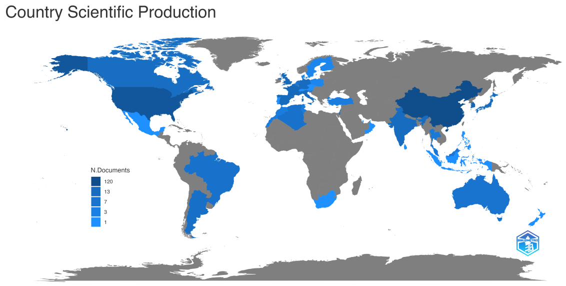

Fig. 3: Country Scientific Production

knowledge and techniques from multiple fields and different areas of knowledge.

Moreover, contributions from various countries demonstrate the global interest in building lexicons. China, France, South Korea, the UK, and the US have emerged as major players in this segment (Figure 3). This international participation undoubtedly underscores the collaborative efforts and cross-cultural exchange of ideas that enrich the progress of this research area.

The state of the art in lexicon construction reflects an active and dynamic research landscape. The variety of topics explored and the participation of researchers from different countries demonstrate the widespread recognition of this research area. As technology advances and new challenges emerge, lexicon construction is expected to evolve, paving the way for innovative approaches and solutions in the field.

Lexicon construction plays an important role in natural language processing and computational linguistics, providing a valuable resource for various applications such as information retrieval, machine translation, and sentiment analysis.

An important aspect of lexicon construction is the automatic generation of a global thesaurus and the paper ”An approach to the automatic construction of global thesauri” [1] shows and approach to generate it automatically. The contribution of this article is to demonstrate the improvement of search performance through the use of suitable clustering algorithms. The authors provide a valuable approach for improving information retrieval tasks by utilizing term discriminant value models.

In the field of multilingual applications, ”Construction of a Chinese–English verb lexicon for machine translation and embedded multilingual applications” [2] focuses on building a Chinese-English verb dictionary. The corpus definition is not explicitly described, but the lexical knowledge extraction technique is described in this article. Using these methods, the authors enable accurate and comprehensive verb translations, contributing to the improvement of machine translation and multilingual applications.

Some of the articles in the literature review indirectly contribute to the desirable properties of lexicons by presenting innovative approaches and construction methods. For example, ”Lexicon Construction for Information Systems” [3] proposes a strategy for automatically identifying symbols in natural language documents. By using natural language processing techniques, the authors improve the extraction of relevant expressions, thereby contributing to the development of richer vocabularies for information systems.

Additionally, sentiment analysis is an active research field and some articles have made significant contributions to this area. For instance, ”A random walk algorithm for automatic construction of domain-oriented sentiment lexicon” [4] presents an adaptive learning framework for building domain-oriented sentiment lexicons. The authors improve lexicon construction through the use of unsupervised active learning algorithms, enabling more accurate sentiment analysis in specific domains.

Lexicon structure is another important aspect addressed in the state of the art. ”Constructing concept lexica with small semantic gaps” [5] proposes a method to analyze and quantify the semantic gap between concepts. The authors use content attribute information and clustering algorithms to identify concepts with small semantic gaps, thereby contributing to the development of a more consistent and comprehensive concept vocabulary. Some works focus on specific areas, for instance ”Construction of the Semantic Lexicon of Dermatology” [6] shows the application of natural language processing technology in building a semantic lexicon for the field of dermatology. This contribution could improve domain-specific analysis and understanding in the field of dermatology.

Another research focuses on the automatic generation of a thesaurus for Chinese documents [7]. The authors present a Chinese keyword extraction algorithm that speeds up the process and achieves comparable term relationships to previous studies. This advancement will improve knowledge organization and information retrieval. Another research presents an adaptive learning framework for creating transliteration lexicons using web sources [8]. The authors used an unsupervised active learning algorithm to improve dictionary construction and enable accurate transliteration in multilingual applications.

In the area of conceptual lexicon construction, Pluempitiwiriyawej and colleagues propose a mechanism to construct complex conceptual vocabularies from existing alphabetic lexicons [9]. By identifying lexical concepts and discovering conceptual relationships through lexical acquisition and clustering, the authors improve the understanding of word relationships and enable the creation of comprehensive conceptual lexicons.

On the natural language processing area Miyamoto and colleagues identified root causes in reporting aviation safety incidents. Using a weakly supervised semantic vocabulary construction approach, the authors improve the identification of influencing factors, thereby improving accident prevention strategies and flight safety [10].

In the field of sentiment analysis, there is research investigating perceptions of emotional states of words in sentences. Poria et. al. propose the concept of 'word affective states' and consider various features and combinations to contribute to mood analysis and emotion recognition [11].

Another article highlights the domain-oriented structure of sentiment lexicons, using sentiment analysis techniques and ontologies to create domain-specific sentiment lexicons to improve sentiment analysis accuracy and domain relevance in applications such as social media monitoring and brand reputation management [12]. Some other research aims at extracting bilingual lexicons from special comparative corpora [13]. The authors use contextual representation strategies and propose smoothing and prediction techniques that improve the reliability of observations of word co-occurrences, facilitating cross-language information retrieval and machine translation.

Another article proposes a strategy to build a domain-specific sentiment lexicon using constrained label propagation [14]. The authors improve the construction of sentiment lexicons by considering dependency information, previous generic lexicons, and applying contextual and morphological constraints to improve the accuracy of sentiment analysis in specific domains.

Related to Arabic sentiment analysis, Al-Ayyoub and colleagues focused on improving lexical-based approaches [15]. The authors described detailed steps to build the key components of their approach and contribute to the advancement of Arabic sentiment analysis through experiments, challenges, and potential improvements.

In the medical field, one research example focuses on terminology extraction from Polish medical texts [16]. The authors process discharge records using linguistic and statistical methods to assess the range and quality of extracted terms and contribute to the development of medical terminology in Polish.

Another study evaluated a pivot-based approach to bilingual lexicon extraction [17]. By comparing different methods of estimating context vectors based on bilingual corpora, the authors evaluate the effectiveness of the approach in two language pairs and explore effective methods for extracting bilingual dictionaries. Finally, Xue and colleagues [18] proposed a method to automatically analyze a sentiment lexicon specifically for microblogs. The authors used the word2vec tool and the microblogging corpus to create an effective sentiment dictionary that was evaluated in a sentiment analysis task and drives sentiment analysis of microblogging data.

These articles present some innovative approaches to lexicon building by addressing specific challenges in different domains and languages, and improving lexicon accuracy and relevance in a variety of applications. These advances pave the way for further research and development in the field of natural language processing.

III. Corpora and lexicons

A corpus is a collection of documents (texts) written in natural language. An ideal corpus should be made up of documents that can be categorized into a single class. The content of documents should only deal with one or more specific topics. Also, documents should have very few (if any) misspellings, and only contain valuable information. In case of opinions, they should not be ambiguous.

A corpus is transformed into a labeled data set if each of their documents have a label assigned. The tag of a document determine its category or class. The process of labeling the documents is carried out manually in many cases, through the review of experts, or directly or indirectly by the authors of the texts themselves. An example of the latter are the opinions on digital platforms, in which users rate products or services and can write opinions. The score is considered the label of the opinion.

Lexicons are lists of terms (words) that are highly representative of a topic. Lexicons are created from the analysis of the documents that belong to a corpus. When the terms of a lexicon have an associated numerical value, they are said to be scored.

Lexicons created for sentiment analysis are extensive, and have demonstrated to be useful to identify the polarity of documents, and to detect hate speech in texts, among other successful applications. However, these lexicons are not accurate in specific contexts.

The main characteristics of the terms that compose an ad-hoc or domain specific lexicon extracted from a corpus (each document assigned to a category) should be the following:

- Some terms in the ad hoc lexicon have special meaning in the specific domain.
- The terms in the lexicon are highly representative of the domain.
- The terms are used frequently in the documents.
- The terms used in the documents belonging to a category, appear very rarely (or does not at all) in the documents of the other categories.
- The lexicon is not too large, i.e., the number of terms is as small as possible.
- Terms in the lexicon are highly discriminative, they are valuable to identify if a document is related to the domain of the corpus, and also the category to which a document belongs to.
- The class distribution of the terms is balanced, i.e., the number of terms is similar for each category. In the case of scored lexicons the scores consider the class imbalance.

In the next subsection, we explain the notation used to refer a corpus, documents and terms in corpus.

A. Notation

The notation used in the methodology is explained below.

$\mathcal{C}$: Corpus, a group of related documents. These documents are in plain text and are labeled, i.e., each document belong to category (class).

d^i: i-th document $\in \mathcal{C}$

${}^{k}d^i$: d^i tagged as belonging to category k (the i-th document in the corpus $\mathcal{C}$, that belongs to category k)

${}^{k}t_j$: j-th term (word) of category k

t_j^i: j-th term (word) of document d^i

${}^{k}t_j^i$: j-th term of document ${}^{k}d^i$ (the j-th term of the i-th document that belongs to the category k)

$f(t)$: frequency of term t

${}^{k}T$: total terms in category k calculated by

$$ {}^{k}T = \sum_{i} \sum_{j} f\left({}^{k}t_j^i\right) \tag{1}$$

${}^{C}T$: total terms in corpus calculated by

$$ {}^{C}T = \sum_{k} \sum_{i} \sum_{j} \left({}^{k}t_j^i\right) \tag{2}$$

${}^{k}N$: Number of documents tagged as belonging to category k

IV. Methodology

The methodology applied on this paper is explained below.

A. Pre-processing

The first step in building a lexicon is to pre-process the corpus documents. Documents (texts) are processed in the usual way in text mining. Stop words, symbols, numbers, punctuation marks, question marks, and exclamation marks are removed. Extra spaces are also removed. The letters are all converted to lower case.

Some contracted words are expanded, for example, from ”ur” to ”$you\ are$”, from ”$2morrow$” to ”$tomorrow$”, etc. For this, the idea shown in [19] was used.

Although it is possible to find the lexemes of the words, we decided not to use this type of transformation, due to the common mistakes made by the SnowBall, Porter and Lancaster algorithms. Lexemes can lead to lexicons with non-representative or non-sense terms .

B. Term features

For the construction of an ad hoc or specific domain lexicon, the characteristics of an ideal lexicon were considered, even when the corpora analyzed do not meet all the desirable characteristics.

The characteristics (features) extracted from each term that appears in the documents of the analyzed corpus try to evaluate if they meet the desirable characteristics to be incorporated into the lexicon. The features extracted to each term are shown in Table I.

V. Experiments and results

A. Corpora analyzed

To test the proposed method, two corpora were used. In both, each document is assigned a label, so they are treated as labeled data sets.

The first data set, Paper reviews, contains most of the texts in Spanish, and some in English. The second data set, SMS SPAM, contains only texts in English. In each of the corpora, each document is assigned to one of two possible categories.

A summary of the data sets is shown in Table II . Paper reviews data set contains less than 400 documents, whereas SMS SPAM data set has more than 5500 documents.

Table III shows the class distribution of documents in the data sets. The presence of class imbalance in both data sets can observed in Table III. For the Paper reviews data set, the ratio between documents of each class is 2:1, while for the SMS SPAM data set it is 5:1.

The class imbalance problem affects not only Machine Learning classification methods, but also lexicon-based methods. This is due to the large number of terms that appear in the majority class, which contrasts with the much fewer number of terms in the minority class. In both approaches, to improve the predictions, some technique must be applied to compensate for the class imbalance.

In order to assess the quality of the lexicon, we use accuracy as a metric. To do this, we apply 5 fold cross-validation. To determine the category to which a document belongs, its terms are extracted. Subsequently, it is identified if these terms appear in the lexicon. For the terms that do appear in the lexicon, the nine features were recovered. The characteristics with the highest numerical value are calculated, assigning them to the corresponding class. In the end, the class with the largest number of assigned terms is the one that is taken as the prediction.

B. Results

The ”Papers reviews” and ”SMS SPAM” corpus were processed as explained above.

For the ”Papers reviews” data set, the 379 documents in the corpus use 31086 words in total. From this amount, 7660 words are used in documents that correspond to accepted papers, and 4973 in documents that correspond to rejected papers. This corpus uses 6102 unique words. The number of terms to which the term-features were calculated was 2517. Of this amount, only 59 were extracted, and added to the lexicon. Only 2% of the terms are considered highly representative to describe the documents.

On the other hand, the ”SMS SPAM” data set contains 5572 documents, which use 61338 words in total. The number of unique words in this corpus is 8147. The number of terms extracted is 1920. In this corpus the 32% of terms are considered representative.

In order to assess the quality of the extracted lexicons, accuracy was used as a performance metric.

The accuracy for the ”Papers reviews” dataset achieved is 81.2% and 84.4% for ”SMS SPAM”.

TABLE I: Term features

Feature	Description
$\mathcal{F}_1(k,j) = \frac{\sum_i f\left({}^k t_j^i\right)}{{}^k T}$	Total repetitions of the term t_j by all documents of class k, divided by the total number of terms in class k
$\mathcal{F}_2(k,j) = \frac{\sum_i f\left({}^k t_j^i\right)}{{}^C T}$	Total occurrences of the term t_j in all documents of class k, divided by the total number of terms in the corpus
$\mathcal{F}_3(k,j) = \sum_i \frac{f\left({}^k t_j^i\right)}{\sum_{jj} f\left({}^k t_{jj}^i\right)}$	Accumulated number of frequency of the term t_j divided by the total terms in the document d^i for each document of class k
$\mathcal{F}_4(k,j) = \frac{\sum_i f\left({}^k t_j^i\right)}{\sum_i 1 \text{ if } {}^k t_j^i \neq 0}$	Total occurrences of the term t_j in all documents of class k divided by the total number of documents in which the term t_j appears
$\mathcal{F}_5(k,j) = \frac{\sum_i f\left({}^k t_j^i\right)}{\bigwedge_{jj}\left(\sum_i 1 \text{ if } {}^k t_{jj}^i \neq 0\right)}$	Total occurrences of the term t_j in all documents of class k divided by the minimum number of occurrences of the term t_j among all documents
$\mathcal{F}_6(k,j) = \frac{\sum_i f\left({}^k t_j^i\right)}{\bigvee_{jj}\left(\sum_i 1 \text{ if } {}^k t_{jj}^i \neq 0\right)}$	Total occurrences of the term t_j in all documents of class k divided by the maximum number of occurrences of the term t_j among all documents
$\mathcal{F}_7(k,j) = \left(\sum_i f\left({}^k t_j^i\right)\right) \cdot \log_2\left({}^C T\right)$	Total repetitions of the term t_j in all documents of class k by the logarithm of the total terms in the corpus
$\mathcal{F}_8(k,j) = \left(\sum_i f\left({}^k t_j^i\right)\right) \cdot \log_2\left(\sum_i f\left({}^k t_j^i\right)\right)$	Total repetitions of the term t_j in all documents of class K by logarithm of the same value.
$\mathcal{F}_9(k,j) = \left(\frac{\sum_i f\left({}^k t_j^i\right)}{{}^k T}\right) \cdot \log_2\left(\frac{{}^k T}{\sum_i f\left({}^k t_j^i\right)}\right)$	$\mathcal{F}_1(k,j)$ times log of $\frac{1}{\mathcal{F}_1}(k,j)$

TABLE II: Data sets

Name	Documents	Description	doi
Paper reviews	379	Paper reviews sent to an international conference mostly in Spanish (some are in English), expressing the reviewer's opinion about the paper.	10.24432/C50G60
SMS SPAM	5572	A collection of SMS spam messages was manually extracted from the Grumbletext Web site.	10.24432/C5CC84

TABLE III: Class distribution

Data set	Documents distribution	Terms distribution
Paper reviews	257 accept	7660 accept
	122 reject	4973 reject
SMS SPAM	4285 ham	12103 ham
	747 spam	4083 spam

TABLE IV: Terms extracted

Data set	Total terms	Terms extracted
Paper reviews	2517	59
SMS SPAM	5975	1920

C. Conclusions

In this paper, we introduced a novel approach for constructing domain-specific or ad-hoc lexicons and thoroughly evaluated its effectiveness on two different corpora. To achieve this, the extraction of nine statistics of each term from the documents that make up a corpus was proposed, and in

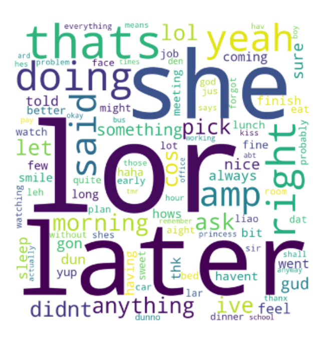

Fig. 4: Summary of lexicon (class ham) constructed from corpus SMS SPAM

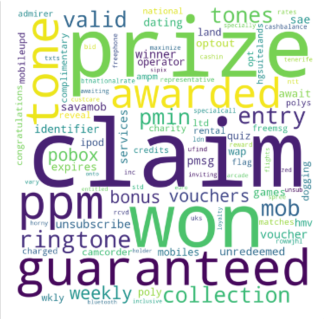

Fig. 5: Summary of lexicon (class spam) constructed from corpus SMS SPAM

addition to considering the categories of the documents. A crucial observation was made during the experiment that class imbalance was present in both datasets. This class imbalance problem is a common challenge in machine learning classification methods and lexical-based approaches because the uneven distribution of terms across classes adversely affects prediction accuracy. To mitigate this problem and improve predictions, we adopted an effective technique of correcting for class imbalance. To assess lexicon quality, we used accuracy as a performance measure and five-way cross-validation to ensure robustness. The proposed method created a domain-specific lexicon considering document categories. These term features were calculated based on their frequency of occurrence relative to the total number of terms in each class and across the corpus. In addition, we analyzed the presence of each term in relation to each class of documents to further improve the accuracy of the lexicon. A particular advantage of this approach is that it can classify new documents without tuning the model, making it efficient and applicable to real-world scenarios. Moreover, the method proved to be robust in processing corpora with imbalanced

References

[1] C. J. Crouch, "An approach to the automatic construction of global thesauri," *Information Processing & Management*, vol. 26, no. 5, pp. 629–640, 1990.

[2] B. J. Dorr, G.-A. Levow, and D. Lin, "Construction of a chinese–english verb lexicon for machine translation and embedded multilingual applications," *Machine Translation*, vol. 17, pp. 99–137, 2002.

[3] M. Sayão and G. R. de Carvalho, "Lexicon construction for information systems," *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 11, no. 36, pp. 35–42, 2007.

[4] S. Tan and Q. Wu, "A random walk algorithm for automatic construction of domain-oriented sentiment lexicon," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12 094–12 100, 2011.

[5] Y. Lu, L. Zhang, J. Liu, and Q. Tian, "Constructing concept lexica with small semantic gaps," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 288–299, 2010.

[6] Y. ZHOU, N. XIANG, R.-X. WANG, Y. LIU, X.-L. QI, and Z.-G. WANG, "Construction of the semantic lexicon of dermatology," in *Proceedings of the Annual Conference of Biomedical Fuzzy Systems Association 28*. Biomedical Fuzzy Systems Association, 2015, pp. 359–362.

[7] Y.-H. Tseng, "Automatic thesaurus generation for chinese documents," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 13, pp. 1130–1138, 2002.

[8] J.-S. Kuo, H. Li, and Y.-K. Yang, "Active learning for constructing transliteration lexicons from the web," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, pp. 126–135, 2008.

[9] C. Pluempitiwiriyawej, N. Cercone, and X. An, "Lexical acquisition and clustering of word senses to conceptual lexicon construction," *Computers & Mathematics with Applications*, vol. 57, no. 9, pp. 1537–1546, 2009.

[10] A. Miyamoto, M. V. Bendarkar, and D. N. Mavris, "Natural language processing of aviation safety reports to identify inefficient operational patterns," *Aerospace*, vol. 9, no. 8, p. 450, 2022.

[11] S. Poria, A. Hussain, and E. Cambria, *Multimodal sentiment analysis*. Springer, 2018.

[12] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 355–363.

[13] E. Morin and E. Prochasson, "Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora," in *Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web*, 2011, pp. 27–34.

[14] S. Huang, Z. Niu, and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, vol. 56, pp. 191–200, 2014.

[15] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of arabic sentiment analysis," *Information processing & management*, vol. 56, no. 2, pp. 320–342, 2019.

[16] M. Marciniak and A. Mykowiecka, "Terminology extraction from medical texts in polish," *Journal of biomedical semantics*, vol. 5, no. 1, pp. 1–14, 2014.

[17] J.-H. Kim, H.-S. Kwon, and H.-W. Seo, "Evaluating a pivot-based approach for bilingual lexicon extraction," *Computational intelligence and neuroscience*, vol. 2015, pp. 45–45, 2015.

[18] B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with word2vec," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 358–363.

[19] E. Osuagwu, "An analysis of chat abbreviations and slangs of the students of the university of port harcourt," *English Linguistics Research*, vol. 9, p. 22, 06 2020.