Scoping natural language processing in Indonesian and Malay for education applications

Zara Maxwell-Smith[1] **Michelle Kohler**[2] **Hanna Suominen**[1,3]
1. The Australian National University / Canberra, ACT, Australia
2. University of South Australia / Adelaide, SA, Australia
3. University of Turku / Turku, Finland
Zara.Maxwell-Smith@anu.edu.au, Michelle.Kohler@unisa.edu.au,
Hanna.Suominen@anu.edu.au

Abstract

Indonesian and Malay are underrepresented in the development of natural language processing (NLP) technologies and available resources are difficult to find. A clear picture of existing work can invigorate and inform how researchers conceptualise worthwhile projects. Using an education sector project to motivate the study, we conducted a wide-ranging overview of Indonesian and Malay human language technologies and corpus work. We charted 657 included studies according to Hirschberg and Manning's 2015 description of NLP, concluding that the field was dominated by exploratory corpus work, machine reading of text gathered from the Internet, and sentiment analysis. In this paper, we identify most published authors and research hubs, and make a number of recommendations to encourage future collaboration and efficiency within NLP in Indonesian and Malay.

1 Introduction

Limited natural language processing (NLP) resources currently available for Indonesian and Malay varieties do not reflect large speaker populations of these languages in Indonesia, Malaysia, and other South-East Asian nations[1]. Difficulties locating resources and existing work hinders progress in the field; it can result in duplicated or unnecessary work, clouding the ability of researchers to formulate useful research questions and study designs. Since Indonesian and Malay varieties are closely related (Sneddon, 2003; Basuki and Antaputra, 2020b), connecting research and technologies developed for either language could provide useful insights and shortcuts for work in the other language.[2]

These challenges restrict the impact that advances in NLP might have in the education sector in Indonesia and Malaysia, and in the teaching of these languages. Ideally, teachers of Indonesian or Malay as a second or foreign language would draw on a wide range of human language technologies, machine learning methods, and corpus linguistics tools to enhance teaching and learning outcomes [3].

As part of a broader project investigating teacher-speech and materials for Indonesian language teaching (Maxwell-Smith et al., 2020; Maxwell-Smith, 2021), the aims of this study were to scope the state of play of existing work in Indonesian and Malay NLP to assist in the formulation of realistic research goals, and to identify useful networks and resources. As such, our study draws on the notion of scoping work as "reconnaisance" (Peters et al., 2015), where the goal is to first determine what range of quantitative and/or qualitative evidence is available on a topic and then to chart, map, or otherwise represent this located evidence visually.

Our research questions were as follows:

1. What language technologies and NLP resources exist for Indonesian/Malay (and therefore for education sector applications)?

2. How do they align with the trends seen more widely in NLP?

We begin by describing our search strategy and methods for charting 657 included studies by their

[1] In 2011, the Indonesian Census recorded 197 million Indonesians as literate in Indonesian (Zein, 2020). In Malaysia, nearly the whole population speak Malay as a first or additional language (Coluzzi, 2017); in 2021, according to the Department of Statistics Malaysia, this was about 32 million people.

[2] As indicated in Lin et al. (2019c) and Nomoto et al. (2018a), some significant differences should caution NLP researchers from regarding the Indonesian and Malay languages as one.

[3] See for example Lee et al. (2020) in journals such as CALL and LLT.

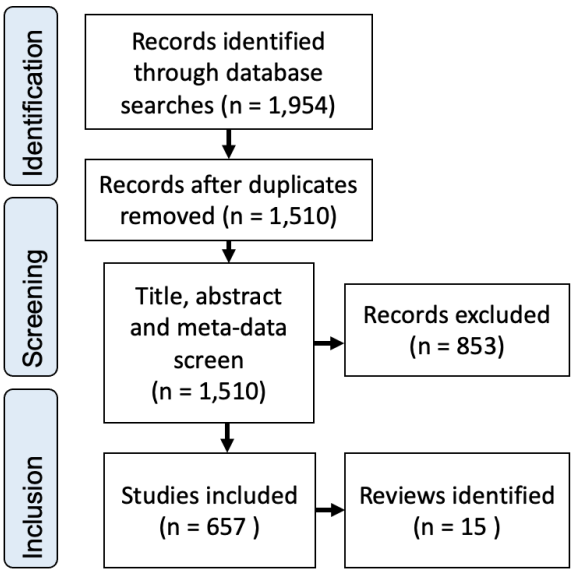

Figure 1: Screening Process and Study Selection

topic and year of publication. We summarize existing literature reviews and notable work, graph most published authors and their affiliated research hubs, and describe recent educational applications of NLP. Finally, we make 7 recommendations to build collaboration and efficiency in Indonesian and Malay NLP, highlighting our contribution to these goals.

2 Methods

In order to capture a broad and rich picture of the recent literature, we applied a simplification of the Systematic Reviews and Meta-Analyses (PRISMA) for scoping reviews (Tricco et al., 2018; Page et al., 2021). Generally, we aimed to provide a descriptive overview and visualization of the reviewed material without detailed critical appraisal of individual studies or synthesis of evidence from different studies (Pham et al., 2014; Peters et al., 2015).

The review engaged with literature from many disciplines, including, but not limited to, Computational Linguistics, Computer Science, Indonesian Language Teaching, and NLP. Extensive consultations with research librarians resulted in a broad search strategy of the databases and terms (Figure 2). Both Indonesian/Malay and English search terms were used to maximize coverage. Additional search terms focused on our interest in Indonesian language teaching were used to mitigate the risk of missing relevant literature.

We experienced significant problems identifying studies with the Association for Computational Linguistics (ACL) as their publisher. In searches via Google Scholar, Scopus, and Proquest; many ACL publications (e.g., Koto et al. (2020a) and Wilie et al. (2020)) were not returned[4]. We then added a direct ACL Anthology search to our database search list. Unexplained behaviour in the ACL Anthology sort by 'Year of Publication' functionality cut results by over 400%, missing, for example, the aforementioned two relevant papers. To identify a reasonable portion of the work presented at ACL events, we extended the set of identified records semi-automatically by manually opening and exporting studies from ACL Anthology searches.

Our inclusion criteria specified that studies be recent (published in 2016 or later), peer-reviewed, written in English, Indonesian or Malay, and relevant to our topic (work about other languages or unrelated to NLP was excluded). Topical relevance was determined by a single reviewer (the first author of this paper) screening the title, abstract, and metadata of each identified study ($n = 1,954$) that was unique ($n = 1,510$) (Figure 1).

Included studies were then classified according to Hirschberg and Manning's 2015 characterization of advances in NLP. We refer to these classifications in brief as: *Broad NLP*; *Machine Reading*; *Machine Translation*; *Spoken Dialogue Systems*; *Speaker State*; and *Social Media*. We added a class — *Statistical Work* — to group work which primarily contributes corpus data or statistical and preprocessing work which stands at the foundation of most NLP.

Two reviewers (the first and last author of this paper) worked as a classification team, thereby assuring the quality of this 'light-touch' manual content analysis (Saldaña, 2016). In total, 80 of the 657 included studies (12.2%) were classified independently by both reviewers, including studies whose classification was perceived as uncertain by the first reviewer, as well as a random selection of further studies to increase confidence in reviewer agreement. Reviewers' classification disagreements were resolved through reference to full-text articles and discussion reaching consensus[5].

While many studies were classified as belonging to more than one 'grouping', the primary class was

[4]Not in the top 100 search results on Google Scholar.

[5]Both reviewers are academic researchers in NLP and teachers. Reviewer agreement of classification was very high. Most often the selection of a primary class was discussed and resolved by looking beyond brief, and at times misleading, information in article abstracts.

Database/Search Engine and search date/details	NLP and Indonesian/Malay keywords	Additional education keywords added to search string
Scopus **ProQuest** **EBSCO** **Limited to peer-review** **21 October 2021**	("natural language processing" OR nlp OR corpus OR corpora OR "computational linguistics" OR "pemrosesan bahasa alami" OR "pengolahan bahasa alami" OR korpus OR korpora OR "linguistik komputasional") AND ("indonesian language" OR "bahasa indonesia" OR malay)	AND ("language teaching" OR "language learning" OR "foreign language" OR bipa OR tifl OR tisol OR "belajar bahasa")
Google **Reduced length due to character limit** **8 July 2021**	("natural language processing" OR corpus OR corpora OR "computational linguistics" OR "pemrosesan bahasa alami" OR "pengolahan bahasa alami" OR korpus OR korpora OR "linguistik komputasional") AND ("indonesian language" OR "bahasa indonesia" OR malay)	("natural language processing" OR corpus OR "computational linguistics" OR "pemrosesan bahasa alami" OR korpus OR "linguistik komputasional") AND (indonesia OR malay) AND ("language teaching" OR "language learning" OR bipa OR "belajar bahasa")
ACL Anthology **25 October 2021**	(indonesian OR malay OR bahasa) AND ("language teaching" OR "language learning" OR "foreign language")	indonesian OR malay OR bahasa

Figure 2: Search Strategy

used in our analysis below. Appendix B is sorted by the second and third classification levels to improve search-ability and provide further information.

A search of titles and abstracts uncovered pre-existing reviews. These reviews were screened in full text and their findings are outlined in our results section to complement the scoping or quantitative map of the field. Literature outside our inclusion criteria which appeared highly relevant was retained separately for full-text review.

Unique names in the raw list of the top 50 most-published authors were manually normalized to prepare a publication-by-author count. Manual identification of name variants for authors with many publications were identified by matching author initials, affiliations, and profiles where possible to create Figure 5. Author affiliations for Figure 6 were taken from the most recent study of a given author included in this overview.

Figure 3: Indonesian and Malay NLP Research in 2016–2021

3 Results

A total of 657 from 1,954 studies met our inclusion criteria. Statistical and corpus work dominated throughout the last 5 years (from 2016 — Figure 3). Studies related to machine reading and sentiment analysis of online text such as news websites (included in 'Speaker States') and social media (sentiment analysis comprises much of our 'Social Media' classification) were popular and showed growth (Figure 4). The largest growth area was in 'Speaker Dialogue Systems' with a relative boom in publishing in 2019 (26 studies).

While our search terms were bilingual, the majority of studies that met our inclusion criteria were written in English or had an English title and abstract. The apparent stagnation of publications in 2020 (Figure 4) with a rebound in the first half of 2021 (which our search covered) could be related to the context of the COVID19 pandemic.

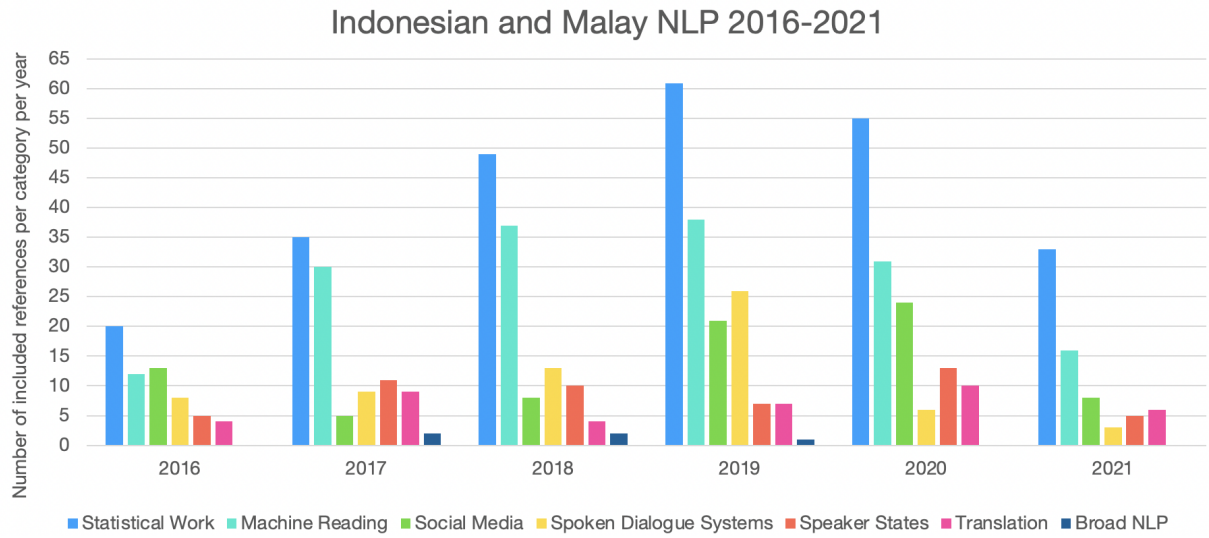

Figure 4: Categorisation of Indonesian and Malay NLP Research in 2016–2021 by Year

However, as "country-specific variables play a significant role" (Abramo et al., 2022) in pandemic publication trends, this is very difficult to determine.

As noted by Hirschberg and Manning in 2015, machine reading research makes use of the vast quantities of text available in the modern world while the mining of text from social media has "revolutionized the amount and types of information available today to NLP researchers". The presence of work in all classifications indicates Indonesian and Malay NLP has advanced from its 2015 state, when Hirschberg and Manning said it had "no such resources or systems available" (2015). To complement the quantitative picture in Figures 3 and 4, we identified existing reviews and provide a summary below. A detailed full-text review of all articles identified in a given classification is needed to fully investigate progress in each respective field.

3.1 Existing Reviews

Of the included studies, 15 were identified as reviews and read in full-text. Lan and Logeswaran (2020) discussed NLP in Indonesian/Malay in general. They did not identify their search methods nor their inclusion criteria, and their reference list had a strong focus on Malaysian research. Their description of statistical work on morphological and lexical analysis, the development of stop word lists, text normalization and named entity recognition concluded that "most researchers have no choice but to resort to compile their own corpus specific to their domain" (Lan and Logeswaran, 2020). According to their discussion, applications of NLP, such as those for machine translation, sentiment analysis, sarcasm and spam detection, as well as text summarization, were hampered by an absence of language-specific tools and resources. For example, they stated that the Jawi Malay script (which is based on Arabic) appeared to be missing characters, lemmatizers for translation seemed to struggle with affixes as they were loaned from English NLP, and sentiment analysis tended to rely on translated sentiment lexicons.

A 16th review article — "An Overview of Natural Language Processing for Indonesian and Malay" by Jiang et al. (2020), written in Mandarin — was identified at the screening stage. It fell outside the scope of this study but we did note that it provided a detailed overview of Indonesian/Malay NLP. The authors characterized the field as "widely distributed, covering stemming, part-of-speech tagging, syntactic analysis, semantic analysis, and other underlying technologies, as well as upper-level applications such as machine translation, spell checking, sentiment analysis, named entity recognition" (Jiang et al., 2020). However, similarly to Lan and Logeswaran (2020), they noted that "the basic resources, open data platforms and open-source language processing tools for these two languages are also lacking, and there are few mature and available text analysis systems" (Jiang et al., 2020).

Statistical or corpus based NLP is important to further the field; however, only 2 reviews had a special focus on corpora and these were specific to Malay (Awang Abu Bakar et al. (2018) and Nasharuddin et al. (2018)). These reviews provided

some insight into Malay resources, and Nasharuddin et al. (2018) suggested document alignment as an avenue to overcome parallel corpus scarcity in cross and bilingual information retrieval, however, a thorough picture of existing corpora was lacking.

A further review by Kassim et al. (2016b) discussed morphology related challenges in stemming tools for Malay as perceived by the authors. However, this review did not fully address the complex steps necessary to uncover lemmas. As later described by Nomoto (2020), what has been "thought of as stemming and lemmatization [· · ·] is in fact 'root'-ing, that is, undoing all morphological processes to get a root". Future work needs to make use of the sort of stem and lemma information in MALINDO Morph to create Indonesian/Malay stemmers.

Machine translation was discussed in a single review of Indonesian translation by Rahutomo et al. (2019)[6], who identified that many researchers had created their own web-crawled parallel corpora. They described a range of techniques used in Indonesian translation, noting that Moses was commonly used and that attention-based approaches were improving neural machine translation. Their list of studies spanned languages: Sundanese, Javanese, Lampung, as well as English, Japanese, and Korean — a very limited list given there are between 652–701 languages in use in Indonesia (Zein, 2020). Thereby translation needs are yet to be met.

Machine reading was the focus of 5 reviews. Gunawan and Amalia (2018) reviewed single document text summarization and identified evaluation methods as a significant concern among 10 papers reporting research into extractive text summarization. They concluded that a text-summary dataset created by experts is needed to advance the field and to calibrate the diverse results reported in the literature.

Looking only at Malay, Mohemad et al. (2020b) suggested relatively poor results across the field. They found summaries were often longer than the original text and Malay anaphora proved difficult to condense, resulting in poor comprehensibility.[7]

In 2021, Widodo et al. remained concerned with evaluation measures in text summarization. Their review of 6 studies found all text summarization work was in extractive summarization — as opposed to abstractive — and that it was dominated by single document summarization of online news. To expand the usability and scope of these tools for Indonesian, they suggested that journal articles should be used as data to support multi-document summarization, as existing summaries of these documents could be used to enhance results.

Malay named entity recognition and classification (NERC) was carefully reviewed by Mohemad et al. (2020a), finding that differences in Malay morphology and textual ambiguities, as well as limitations on corpora and annotated data, are difficult challenges affecting both rule-based and machine learning methods. In addition, they found that the "majority of the systems developed [were] based on manually predefined dictionaries by a human" (Mohemad et al., 2020a) and that deep learning methods were yet to be studied with Malay NERC.

All 4 reviews of sentiment analysis were primarily concerned with social media in the Malaysian context (see both 'Speaker States' and 'Social Media' in Appendix B). Abdullah et al. (2017) found hybrid approaches of lexicon based and supervised machine learning were most common, while Handayani et al. (2018) added a more detailed discussion of techniques and datasets found in 10 carefully selected studies. Abu Bakar et al. (2020) foregrounded the 'noise' of social media data to confront the more complex language often found on the Internet. Abdullah and Rusli (2021) pushed this further, examining literature on multilingual sentiment analysis to inform the development of sentiment analysis for the Malaysian social media context, which they described as characterized by the multilingual use of English, Malay, and Chinese.

No reviews of spoken dialogue systems, such as automated speech recognition (ASR) or Text To Speech (TTS) toolkits (which are typically considered later-generation NLP), were found. This is not surprising given text-based NLP (e.g., machine reading) dominated the research agenda (Figure 4).

3.2 Notable Work Responding to the Lack of Data and Evaluation Methods, and Other Recent Contributions

Common to all reviews was a scarcity of freely available NLP resources, and subsequently the creation of custom datasets, loaned preprocessing tools from NLP in English, and difficulties in benchmarking performance without reliable eval-

[6] see also Septarina et al. (2019)

[7] Providing a brief reference to Malay language corpora, Omar et al. (2021) outlined advances in text summarization techniques.

uation techniques and reference datasets. In this context, we note the growing use of zero and few shot methods which are supported by pipelines such as HuggingFace[8]. We also note four projects and respective papers that develop benchmarking and open access corpora for:

- language modelling; Indonesian Language Evaluation Montage (IndoLEM) and Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT);
- Indonesian Natural Language Understanding (IndoNLU); as well as
- Indonesian Natural Language Generation (IndoNLG) and Indonesian Natural Language Inference (IndoNLI).

Recently available, but not included in our study as it was published after our analysis was complete, Aji et al. (2022) provide a detailed discussion of NLP for the 700+ languages spoken in Indonesia. They outline challenges for NLP in Indonesia, describing limited resources, language diversity, orthography variation, and societal challenges such as the poor distribution of technology and education across Indonesia.

3.3 Highly Active Researchers and Research Hubs

Highly active researchers in the field are identified in Figure 5. While we made substantial efforts to ensure author publications were grouped accurately, name variations appeared to be prevalent in this field. We concentrated our efforts on normalizing the raw list of the top 50 authors. In contrast to broader trends (Mohammad, 2020), we note that 7 of the 11 authors in Figure 5 are female, though some are not first authors on many papers.

Research hubs in Indonesia and Malaysia are illustrated in Figure 6. The affiliations of the 25 authors with the most publications were used to identify these hubs. All affiliated universities listed by these 25 authors (as indicated in their most recent publication which met our inclusion criteria) were in either Malaysia or Indonesia. While there was an even spread between the two countries, overall Malaysia dominated with a ratio of 14:11 affiliations in Malaysia and Indonesia, respectively.

[8] See, for example, Cahya Wirawan's pre-trained Wikipedia model.

3.4 Education Specific Studies

The number of education specific studies was 41 (see Appendix A), based on the title and abstract. Of these 41 studies, 15 focused on assessment, with an emphasis on expediting and improving the efficiency of grading and providing feedback. Earlier studies (2016–2018) tended to focus on word-level error correction and short-answer grading while later studies (2019–2021) seemed to address whole of text evaluation, assessment task design (particularly questioning techniques), and providing feedback. Within the limited time frame of our study, we tentatively noted a shift from micro language level applications and their intrinsic evaluations to more macro, holistic language use applications that proceed to extrinsic or broader NLP evaluations.

A portion of education studies considered teaching practices and teacher training. Generally the studies reflected the design of our search-terms to target Indonesian language teaching; 10 papers were geared towards using NLP to improve the teaching and learning of Indonesian/Malay for non-background language learners. Bahasa Indonesia bagi Penutur Asing (BIPA — *Indonesian language for Foreign Learners*) and the Malay equivalent were discussed separately. Two studies were related to using NLP to improve the training of teachers of Indonesian as a foreign language. Another 2 studies focused on NLP for improving the teaching of translation for local students (i.e., Indonesian background speakers). Beyond studies from a language teaching setting, a further 6 studies related to instruction in general or other areas of the curriculum such as Mathematics, study skills, and values education.

Overall the body of work indicates a need for greater resourcing generally, and greater resourcing in education, with a shift towards more sophisticated language concerns and potential uses for these methods. Pleasingly, researchers identified for a high number of publications in Figure 5, such as Amalia, A[9] were also identified among those developing NLP for education (Amalia et al., 2019a), indicating high profile NLP researchers are invested in education sector applications.

4 Discussion

This study sheds significant light on the state of play and progress of Indonesian and Malay NLP.

[9] (see also name variant: *Amalia, Amalia* and Google Scholar profile *Amalia, Mahdi*)

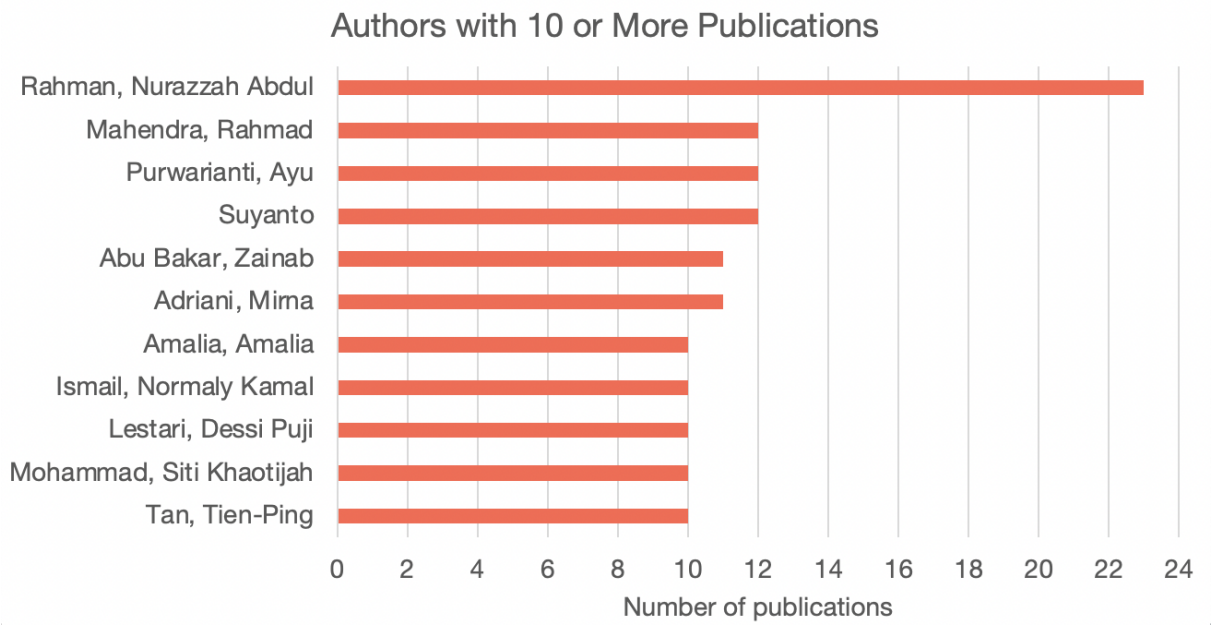

Figure 5: Authors with 10 or More Publications

Figure 6: Author Affiliation of the Top 25 Authors with Most Publications

We make the following 7 recommendations and outline our related contributions to encourage collaboration and support appropriate investment in the development of Indonesian and Malay NLP, and its application in education.

4.1 Recommendations

1. Broader adoption of best-practices for findable, open, sustainable, and future-proof data. All 15 reviews raised the difficulty of locating data, tools, and existing work as a significant problem. Our study responds to this problem and is unique in that Appendix B offers a catalogue of recent work, identifies a variety of openly shared datasets and tools, and locates research hubs. Our results complement reviews such as Lan and Logeswaran (2020) and Jiang et al. (2020) by providing an empirical picture of published research on Indonesian and Malay NLP. While our study was not specifically focused on corpora like Awang Abu Bakar et al. (2018), over 50 of the papers listed at the top of 'Statistical Work' in Appendix B point to

datasets which may be useful to future projects.

Useful advice to implement the FAIR Principles (Findable, Accessible, Interoperable, Reusable) and further advice for individual researchers/teams to develop metadata, choose file formats, and think beyond their immediate plans for linguistic data is set out by Janda (2022) and Mattern (2022), respectively. Funding requirements that encourage data reuse (given ethics restraints) and which provide financial support for adequate digital stewardship are recommended.

2. Expanded evaluation methods and datasets, and negative result publication

As identified by Gunawan and Amalia (2018) and Widodo et al. (2021), evaluation of Indonesian and Malay NLP is poorly supported due to a lack of clear methods and few reference datasets. Open, accessible data from studies can assist with this, but to encourage further efficiency in the field, we also recommend authors consider the publication of 'unsuccessful' experiments in venues such as Workshop on Insights from Negative Results in NLP and organize or participate in evaluation challenges (a.k.a. shared tasks) and their workshops, for example, as part of ACL conferences.

3. Collaboration and connection between Indonesian and Malay NLP research and projects to speed development and allow cross-fertilisation

Many of the reviews discussed in our results focused on Malaysian research specifically looking at Malay NLP. By using both Indonesian and Malay in our search terms we connect these reviews to the work of Indonesian authors. To illustrate, Handayani et al. (2018), Abu Bakar et al. (2020), and Abdullah et al. (2021) examine one study which included Malay in the application of multilingual sentiment analysis. Our findings connect this work in Malay to more recent work by Tho et al. (2021), who looked at Indonesian and Javanese code-mixed sentiment analysis.

Indonesian and Malay are closely related (Sneddon, 2003). With a caveat that corpus metadata must clearly describe which languages are present, and that projects must clearly state how Indonesian and Malay are used in training data, we recommend future work seek out synergies that could be leveraged by using both languages.

4. Flexible author name formats and consistent author name use

Many authors of studies we included had only one name, but appeared to double this name in some publications but not others, perhaps to suit forms built with an (eurocentric) expectation of family names. Similarly, many authors with lengthy names used various forms.

We recommend publishers adapt their submission forms to accommodate diverse name traditions and support existing unique author identifiers (e.g., the ORCID system). We also recommend authors choose a publication name and use it as consistently as possible to increase the findability of their work.

5. Investment in spoken language data and transcription protocols

Our findings indicate only modest developments of 'later-generation' NLP in Indonesian and Malay. Significant investment is needed to give users of these languages access to a broader gamut of NLP applications, including applications in the education sector.

In this space it is also essential to recognise differences in the actual usage of these languages in real-life, spoken situations. Transcription which records code-mixed and often diglossic use of spoken varieties of Indonesian and Malay in a machine readable format needs to be investigated and scrutinised (Maxwell-Smith et al., 2020).

6. Investment in other languages of Indonesia and Malaysia

As a necessary endeavour for equitable access to advances in NLP for speakers, and to limit the further endangerment of many languages as a consequence of the expansion of Indonesian (Zein, 2020), we recommend simultaneous investment in other languages of Indonesia and Malaysia. Linked to Recommendation 5, to reflect actual usage and to allow NLP to be useful in real-world contexts where code-mixing is the norm, investment in other languages is also likely crucial.

7. Education and NLP researchers should consider the use of datasets by researchers outside their field

A research project such as ours, investigating teacher-speech and teaching materials for Indonesian language teaching (Maxwell-Smith et al., 2020) should take advantage of human language technologies. This article contributes a language-specific characterization of the field which will help scope future projects.

Education researchers should be aware that computational methods such as data normalization scripts and stemming tools are yet to be fully de-

veloped for their use with the Indonesian language. If working with spoken language, ASR toolkits for working on low resource languages are suitable for consideration but may require significant investment of time and training before they are capable of managing complex code-switching behaviours common in education settings.

For education researchers and teachers to use NLP resources they need clear information about the profile of language/s in corpora and also about what data has been used when training models/tools. This allows proper assessment of the cultural and political suitability of NLP resources.

Education researchers also need to consider the possible use of datasets by researchers outside their field. Ensuring data they collect is recorded in 'future-proof' formats and prepared with consideration of the FAIR principles (see Recommendation 1 — Janda (2022) and Mattern (2022)) is an investment which encourages NLP applications specifically built for or amenable to education settings.

4.2 Limitations and Future Work

With regard to the limitations of this study, we employed a 'light-touch', subjective coding method with a discrete set of class labels to scope relevant literature; we did not undertake the act of synthesis (Peters et al., 2015). We screened only the title and abstract for the vast majority of references. Unintentional misinterpretation could have taken place. Our analysis provides an approximate area within NLP for each reference to assist researchers studying an NLP application or use-case. Researchers interested in a particular algorithm (e.g., Random Forest Decision Trees or Transformers), or the use of a particular performance indicator (e.g., F1 or Word Error Rate), might not find our work as useful, but we encourage them to scan related categories in Appendix A for work relevant to their interest. Most references were labelled as belonging to multiple categories, with the identification of the first category an educated but ultimately subjective decision. Reading every study carefully beyond title and abstract was beyond the scope of this study.

Future studies to target Indonesian and Malay language publications may identify further literature on Indonesian and Malay NLP. Unfortunately, our initial searches through databases such as the University of Indonesia's Research Portal, produced varied results, with a large proportion of returned studies not necessarily having been as rigorously peer-reviewed (encompassing for example many 'skripsi' or honours dissertations). Indonesia's national library service OneSearch has grown dramatically, and with over 3748 libraries affiliated in February 2022, it should also be included in future reviews of Indonesian and Malay NLP.

Since we conducted our review, Aji et al. (2022) have proposed potential research directions in the Indonesian context such as data-efficient and compute-efficient NLP. Given the low number of studies in our *Speaker Dialogue Systems* class, we support their call for "NLP Beyond Text". The 'superglossic' translanguaging practices of Indonesia (Zein, 2020), and language classrooms (Maxwell-Smith et al., 2020), correspond with their call for "Robustness to Code-mixing and Non-Standard Orthography". Applications of Indonesian NLP necessitate involvement with other languages of Indonesia and inevitably impact many at-risk languages. There is an ethical obligation for "careful assessment of individual usage scenarios of language technology, so they are implemented for the good of the local population" (Aji et al., 2022).

5 Conclusion

Overall, this scoping study provides a baseline picture of Indonesian and Malay NLP. It shows an emerging research community engaged with the wide range of NLP advances identified in 2015 by Hirschberg and Manning. Researchers in the field continue to experience difficulties in benchmarking performance without reliable evaluation techniques and reference datasets, re-engineering of loaned preprocessing tools from English NLP, and thankless tasks such as the creation of custom datasets and resources. NLP applications in education are limited, as are tools for language which is not in text format. Our results highlight the importance of creating and releasing well-described and maintained resources openly and fostering collaboration. IndoLEM-IndoBERT, IndoNLU, IndoNLG, and IndoNLI are notable releases that are already helping to orientate researchers and future projects using Indonesian and Malay NLP.

Acknowledgements

We are grateful to Murray Hall, Chenchen Xu, and Rebecca Barber for their technical, translation, and search strategy assistance, respectively. We also thank the anonymous reviewers for their helpful feedback.

References

Muhammad Aasim Asyafi'Ie bin Ahmad, Mokhtar bin Harun, Puspa Inayat binti Khalid, Mohd Ibrahim Shapiai, Md. Najib bin Ibrahi, and Siti Zaleha Abdul Hamid. 2017. Comparison of the themes of Malaysian Friday sermons between the year 2010 and 2015. *Indonesian Journal of Electrical Engineering and Computer Science*, 6(1):212–218.

Nurazzah Abd Rahman, Nursyahidah Alias, Nomaly Kamal Ismail, Zulhilmi Bin Mohamed Nor, and Muhammad Nazir bin Alias. 2016. An identification of authentic narrator's name features in Malay hadith texts. *IEEE Conference on Open Systems, ICOS 2015*, pages 79–84.

Imran Ho Abdullah, Anis Nadiah Che Abdul Rahman, and Azhar Jaludin. 2021. The development of the Malaysian Hansard Corpus: A corpus of parliamentary debates 1959-2020. *Jurnal Linguistik*, 25(1).

Nur Atiqah Sia Abdullah and Nur Ida Aniza Rusli. 2021. Multilingual sentiment analysis: A systematic literature review. *Pertanika Journal of Science and Technology*, 29(1):445–470.

Nur Atiqah Sia Abdullah, Nurul Iman Shaari, and Abd Rasul Abd Rahman. 2017. Review on sentiment analysis approaches for social media data. *Journal of Engineering and Applied Sciences*, 12(3):462–467.

Enid Zureen Zainal Abidin, Nik Farhan Mustapha, Normaliza Abd Rahim, and Syed Nurulakla Syed Abdullah. 2020. Translation of idioms from Arabic into Malay via Google Translate: What needs to be done? *GEMA Online Journal of Language Studies*, 20(3):156–180.

Zaenal Abidin and Permata Permata. 2021. Pengaruh penambahan korpus paralel pada mesin penerjemah statistik Bahasa Indonesian ke Bahasa Lampung Dialek Nyo. *Jurnal Teknoinfo*, 15(1):13–19.

Zaenal Abidin, Permata Permata, I. Ahmad, and Rusliyawati. 2021. Effect of mono corpus quantity on statistical machine translation Indonesian-Lampung dialect of Nyo. *3rd International Conference on Applied Sciences Mathematics and Informatics, ICASMI 2020*, 1751.

Achmad Fatchuttamam Abka. 2017. Evaluating the use of word embeddings for part-of-speech tagging in Bahasa Indonesia. *2016 International Conference on Computer, Control, Informatics and its Applications, IC3INA 2016*, pages 209–214.

Giovanni Abramo, Ciriaco Andrea D'Angelo, and Ida Mele. 2022. Impact of Covid-19 on research output by gender across countries. *Scientometrics*.

M. Abu, A. Amir, N. A. H. Zahri, and R. Ngadiran. 2020. Voice-based Malay commands recognition by using audio fingerprint method for smart house applications. *IOP Conference Series. Materials Science and Engineering*, 767(1).

Muhammad Fakhrur Razi Abu Bakar, Norisma Idris, Liyana Shuib, and Norazlina Khamis. 2020. Sentiment analysis of noisy Malay text: State of art, challenges and future work. *IEEE Access*, 8:24687–24696.

Muhammad Yuslan Abu Bakar, Adiwijaya, and Said Al Faraby. 2019a. Multi-label topic classification of hadith of Bukhari (Indonesian language translation) using information gain and backpropagation neural network. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 344–350.

Normi Sham Awang Abu Bakar, Ros Aziehan Rahmat, and Umar Faruq Othman. 2019b. Polarity classification tool for sentiment analysis in Malay language. *IAES International Journal of Artificial Intelligence*, 8(3):258–263.

Rike Adelia, Suyanto Suyanto, and Untari Novia Wisesty. 2019. Indonesian abstractive text summarization using bidirectional gated recurrent unit. *4th International Conference on Computer Science and Computational Intelligence, ICCSCI 2019*, 157:581–588.

Rifki Adhitama, Retno Kusumaningrum, and Rahmat Gernowo. 2017. Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology. *1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018-January:247–251.

Ryan Adipradana, Bagas Pradipabista Nayoga, Ryan Suryadi, and Derwin Suhartono. 2021. Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings. *Bulletin of Electrical Engineering and Informatics*, 10(4):2130–2136.

Dwi Intan Af'idah, Retno Kusumaningrum, and Bayu Surarso. 2020. Long short term memory convolutional neural network for Indonesian sentiment analysis towards touristic destination reviews. *2020 International Seminar on Application for Technology of Information and Communication, iSemantic 2020*, pages 630–637.

Afiyati, Edi Winarko, and Anis Cherid. 2018. Recognizing the sarcastic statement on WhatsApp Group with Indonesian language text. *2017 International Conference on Broadband Communication, Wireless Sensors and Powering, BCWSP 2017*, 2018-January:1–6.

Zaaba Ahmad, Syaheerah Lebai Lutfi, Albin Lemuel Kushan, Mohamad Hafiz Khairuddin, Anwar Farhan Zolkeplay, Mohammad Hafidz Rahmat, and Mohd Taufik Mishan. 2017. Construction of the Malay language psychometric properties using LIWC from Facebook statuses. *Advanced Science Letters*, 23(8):7911–7914.

Mohd Zakree Ahmad Nazri, Tri Basuki Kurniawan, Abdul Razak Hamdan, Salwani Abdullah, and Mohammed Azlan Mis. 2018. Taxonomy development from Malay text using firefly bisection algorithm. *GEMA Online Journal of Language Studies*, 18(2):182–201.

Noor Bazilah Ahmat Baseri, Juhaida Abu Bakar, Azizah Ahmad, Hawa Jafferi, and Muhammad Faiz Zamri. 2020. SMVS: A web-based application for graphical visualization of Malay text corpus. *10th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2020*, pages 30–35.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia.

Try Ajitiono and Yani Widyani. 2017. Indonesian essay grading module using natural language processing. *3rd International Conference on Data and Software Engineering, ICoDSE 2016*.

Herley Shaori Al-Ash and Wahyu Catur Wibowo. 2018. Fake news identification characteristics using named entity recognition and phrase detection. *10th International Conference on Information Technology and Electrical Engineering, ICITEE 2018*, pages 12–17.

Tareq Al-Moslmi, Nazlia Omar, Mohammed Albared, and Ade Alshabi. 2017. Enhanced Malay sentiment analysis with an ensemble classification machine learning approach. *Journal of Engineering and Applied Sciences*, 12(20):5226–5232.

Ahmed Al-Saffar, Suryanti Awang, Hai Tao, Nazlia Omar, Wafaa Al-Saiagh, and Mohammed Al-bared. 2018. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS ONE*, 13(4):e0194852.

Andry Alamsyah, Muhammad Fadhli Rachman, Cindy Septiani Hudaya, Rimba Pratama Putra, Aulia Ichsan Rifkyano, and Fivi Nurwianti. 2019. A progress on the personality measurement model using ontology based on social media text. *4th International Conference on Information Management and Technology, ICIMTech 2019*, pages 581–586.

Andry Alamsyah, Sri Widiyanesti, Rizqy Dwi Putra, and Puspita Kencana Sari. 2020. Personality measurement design for ontology based platform using social media text. *Advances in Science, Technology and Engineering Systems*, 5(3):100–107.

Ika Alfina, Ruli Manurung, and Mohamad Ivan Fanany. 2017. DBpedia entities expansion in automatically building dataset for Indonesian NER. *8th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, pages 335–340.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2018. Hate speech detection in the Indonesian language: A dataset and preliminary study. *9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018-January:233–237.

Rayner Alfred, Leow Ching Leong, and Joe Henry Obit. 2017. An evolutionary-based term reduction approach to bilingual clustering of Malay-English corpora. *Proceedings of the International Conference on Advances in Information and Communication Technology, ICTA 2016*, 538 AISC:132–141.

Rayner Alfred, Leow Jia Ren, and Joe Henry Obit. 2016. Assessing factors that influence the performances of automated topic selection for Malay articles. *2nd International Conference on Soft Computing in Data Science, SCDS 2016*, 652:300–309.

Nursyahidah Alias, Nurazzah Abd Rahman, Normaly Kamal Ismail, Zulhilmi Mohamed Nor, and Muhammad Nazir Alias. 2017a. Graph-based text representation for Malay translated hadith text. *3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016*, pages 60–66.

Suraya Alias, Siti Khaotijah Mohammad, Keng Hoon Gan, and Tan Tien Ping. 2018a. MYTextSum: A Malay text summarizer model using a constrained pattern-growth sentence compression technique. *4th International Conference on Computational Science and Technology, ICCST17*, 488:141–150.

Suraya Alias, Siti Khaotijah Mohammad, Gan Keng Hoon, and Tan Tien Ping. 2016. A Malay text corpus analysis for sentence compression using pattern-growth method. *Jurnal Teknologi*, 78(8):197–206.

Suraya Alias, Siti Khaotijah Mohammad, Gan Keng Hoon, and Tan Tien Ping. 2017b. A Malay text summarizer using pattern-growth method with sentence compression rules. *3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016*, pages 7–12.

Suraya Alias, Siti Khaotijah Mohammad, Gan Keng Hoon, and Tan Tien Ping. 2017c. Extract, compress and summarize – An experiment using Malay news article. *Advanced Science Letters*, 23(5):4336–4340.

Suraya Alias, Siti Khaotijah Mohammad, Gan Keng Hoon, and Tan Tien Ping. 2018b. A text representation model using Sequential Pattern-Growth method. *Pattern Analysis and Applications*, 21(1):233–247.

Suraya Alias, Siti Khaotijah Mohammad, Gan Keng Hoon, and Mohd Shamrie Sainin. 2018c. Understanding human sentence compression pattern for Malay text summarizer. *4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences, CAMP 2018*, pages 42–47.

Suraya Alias, Mohd Shamrie Sainin, and Siti Khaotijah Mohammad. 2020. Bilingual extractive text summarization model using textual pattern constraints. *GEMA Online Journal of Language Studies*, 20(3):70–95.

Suraya Alias, Mohd Shamrie Sainin, and Siti Khaotijah Mohammad. 2021. A syntactic-based sentence validation technique for Malay text summarizer. *Journal of Information and Communication Technology*, 20(3):329–352.

A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia. 2019a. Automated Bahasa Indonesia essay evaluation with latent semantic analysis. *Journal of Physics: Conference Series*, 1235(1).

Amalia Amalia, Opim Salim Sitompul, Erna Budhiarti Nababan, Maya Silvi Lydia, and Nadia Rahmatunnisa. 2019b. Bahasa Indonesia text corpus generation using web corpora approaches. *Journal of Theoretical and Applied Information Technology*, 97(24):3810–3821.

Amalia Amalia, Opim Salim Sitompul, Erna Budhiarti Nababan, and Teddy Mantoro. 2020a. A comparison study of document clustering using DOC2VEC versus TFIDF combined with LSA for small corpora. *Journal of Theoretical and Applied Information Technology*, 98(17):3644–3657.

Amalia Amalia, Opim Salim Sitompul, Erna Budhiarti Nababan, and Teddy Mantoro. 2020b. An efficient text classification using fasttext for Bahasa Indonesia documents classification. *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2020*, pages 69–75.

Rizkiana Amalia, Moch Arif Bijaksana, and Dhinta Darmantoro. 2018. Negation handling in sentiment classification using rule-based adapted from Indonesian language syntactic for Indonesian text in Twitter. *International Conference on Data and Information Science 2017, ICoDIS 2017*, 971.

Andi Tenri Ampa, Muhammad D Basri, and Sri Ramdayani. 2019. A morphophonemic analysis on the affixation in the Indonesian Language. *International Journal of Scientific and Technology Research*, 8(7):267–273.

Ahmad Zuli Amrullah, Rudy Hartanto, and I Wayan Mustika. 2017. A comparison of different part-of-speech tagging technique for text in Bahasa Indonesia. *7th International Annual Engineering Seminar, InAES 2017*.

Ratna Anak Agung Putri, Purnamasari Prima Dewi, Boma Anantasatya Adhi, F. Astha Ekadiyanto, Muhammad Salman, Mardiyah Mardiyah, and Winata Darien Jonathan. 2017. Cross-language plagiarism detection system using latent semantic analysis and learning vector quantization. *Algorithms*, 10(2):69.

Muhammad Bagus Andra and Tsuyoshi Usagawa. 2020. Automatic transcription and captioning system for Bahasa Indonesia in multi-speaker environment. *5th International Conference on Intelligent Informatics and Biomedical Sciences, ICIIBMS 2020*, pages 51–56.

Muhammad Bagus Andra and Tsuyoshi Usagawa. 2021. Improved transcription and speaker identification system for concurrent speech in Bahasa Indonesia using recurrent neural network. *IEEE Access*, 9:70758–70774.

Vincent Andreas, Alexander Agung Santoso Gunawan, and Widodo Budiharto. 2019. Anita: Intelligent humanoid robot with self-learning capability using Indonesian language. *4th Asia-Pacific Conference on Intelligent Robot Systems, ACIRS 2019*, pages 144–147.

Miftah Andriansyah, Antonius Irianto Sukowati, Marshal Samos, Imam Purwanto, Ali Akbar, and Muhammad Subali. 2018. Developing Indonesian corpus of pornography using simple nlp-text mining (NTM) approach to support government anti-pornography program. *2nd International Conference on Informatics and Computing, ICIC 2017*, 2018-January:1–4.

Sandhya Aneja, Siti Nur Afikah Bte Abdul Mazid, and Nagender Aneja. 2020. Neural machine translation model for university email application. *2nd Symposium on Signal Processing Systems, SSPS 2020*, pages 74–79.

Dina Anggraini, Achmad Benny Mutiara, Tb. Maulana Kusuma, and Lily Wulandari. 2018. Algorithm for simple sentence identification in Bahasa Indonesia. *3rd International Conference on Informatics and Computing, ICIC 2018*.

Sarudin Anida, Redzwan Husna Faredza Mohamed, Zulkifli Osman, Shah Raja Noor Farah Azura Raja Ma'amor, and Albakri Intan Safinas Mohd Ariff. 2019. Menangani kekaburan kemahiran prosedur dan terminologi awal matematik: Pendekatan leksis berdasarkan teori prosodi semantik. *Malaysian Journal of Learning and Instruction*, 16(2):255–294.

Laksmi Anindyati, Ayu Purwarianti, and Ade Nursanti. 2019. Optimizing deep learning for detection cyberbullying text in Indonesian language. *2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*.

Ani Anisyah, Tricya Esterina Widagdo, and Fazat Nur Nur Azizah. 2019. Natural language interface to database (NLIDB) for decision support queries. *2019 International Conference on Data and Software Engineering, ICoDSE 2019*.

Lalitia Ansari and Totok Suhardijanto. 2019. Where is the head positioned in Indonesian language?: A corpus study of head directionality from a dependency perspective. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 171–177.

Nurfarihah Apandi and Nursuriati Jamil. 2017. An analysis of Malay language emotional speech corpus for emotion recognition system. *2016 IEEE Industrial Electronics and Applications Conference, IEACon 2016*, pages 225–231.

William Aprilius, Seng Hansun, and Dennis Gunawan. 2017. Entity annotation WordPress plugin using TAGME technology. *Telkomnika (Telecommunication Computing Electronics and Control)*, 15(1):486–493.

Winda Widya Ariestya, Ida Astuti, and I Made Wiryana. 2018. Preprocessing for crawler of short message social media. *3rd International Conference on Informatics and Computing, ICIC 2018*.

Siti Noor Allia Noor Ariffin and Sabrina Tiun. 2018. Part-of-speech tagger for Malay social media texts. *GEMA Online Journal of Language Studies*, 18(4):124–142.

Siti Noor Allia Noor Ariffin and Sabrina Tiun. 2020. Rule-based text normalization for Malay social media texts. *International Journal of Advanced Computer Science and Applications*, 11(10):156–162.

Y. Arifin, S. M. Isa, L. A. Wulandhari, and E. Abdurachman. 2018. Plagiarism detection for Indonesian language using winnowing with parallel processing. *Journal of Physics: Conference Series*, 978(1).

Bayu Aryoyudanta, Teguh Bharata Adji, and Indriana Hidayah. 2017. Semi-supervised learning approach for Indonesian Named Entity Recognition (NER) using co-training algorithm. *2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016*, pages 7–12.

Siti Azirah Asmai, Muhammad Sharilazlan Salleh, Halizah Basiron, and Sabrina Ahmad. 2018. An enhanced Malay named entity recognition using combination approach for crime textual data analysis. *International Journal of Advanced Computer Science and Applications*, 9(9):474–483.

Asniar and B. R. Aditya. 2017. A framework for sentiment analysis implementation of Indonesian language tweet on Twitter. *Journal of Physics: Conference Series*, 801(1).

Atqia Aulia, Dewi Khairani, Rizal Broer Bahaweres, and Nashrul Hakiem. 2017a. WatsaQ: Repository of al hadith in Bahasa (case study: Hadith Bukhari). *4th International Conference on Electrical Engineering, Computer Science and Informatics, EECSI 2017*, 2017-December.

Atqia Aulia, Dewi Khairani, and Nashrul Hakiem. 2017b. Development of a retrieval system for al hadith in Bahasa (case study: Hadith Bukhari). *5th International Conference on Cyber and IT Service Management, CITSM 2017*.

Indra Aulia and Ari Moesriami Barmawi. 2016. An automatic health surveillance chart interpretation system based on Indonesian language. *International Conference on Advanced Computer Science and Information Systems, ICACSIS 2015*, pages 163–170.

Indra Aulia, Ainul Hizriadi, Seniman, and Muhibuddin. 2020. Preliminary research design on sensor data gathering for air quality text generation. *4th International Conference on Computing and Applied Informatics 2019, ICCAI 2019*, 1566.

Normi Sham Awang Abu Bakar, Hamwira Yaacob, Dini Handayani, and Mustafa Ali Abuzaraida. 2018. Malay Online Virtual Integrated Corpus (MOVIC): A systematic review. *2018 International Conference on Information and Communication Technology for the Muslim World, ICT4M 2018*, pages 243–248.

Media Anugerah Ayu, Teddy Mantoro, and Jelita Asian. 2018. Quality translation enhancement using sequence knowledge and pruning in statistical machine translation. *Telkomnika (Telecommunication Computing Electronics and Control)*, 16(2):718–727.

Media Anugerah Ayu, Sony Surya Wijaya, and Teddy Mantoro. 2019. An automatic lexicon generation for Indonesian news sentiment analysis: A case on governor elections in Indonesia. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(3):1555–1561.

Indira Suri Azarine, Moch Arif Bijaksana, and Ibnu Asror. 2019. Named entity recognition on Indonesian tweets using hidden markov model. *7th International Conference on Information and Communication Technology, ICoICT 2019*.

Shaari Azianura Hani, Kamaluddin Mohammad Rahim, Fauzi Wan Fariza Paizi, and Mohd Masnizah. 2019. Online-dating romance scam in Malaysia: An analysis of online conversations between scammers and victims. *GEMA Online Journal of Language Studies*, 19(1):97–115.

Jamaluddin Aziz. 2019. Exploring gender issues associated with wanita/woman and perempuan/woman in Malaysian parliamentary debates: A culturomic approach. *GEMA Online Journal of Language Studies*, 19(4):278–303.

Nadhia Salsabila Azzahra, Muhammad Okky Ibrohim, Junaedi Fahmi, Bagus Fajar Apriyanto, and Oskar Riandi. 2020. Developing name entity recognition for structured and unstructured text formatting dataset. *5th International Conference on Informatics and Computing, ICIC 2020*.

Juhaida Abu Bakar, Khairuddin Omar, Mohammad Faidzul Nasrudin, and Mohd Zamri Murah. 2016. NUWT: Jawi-specific buckwalter corpus for Malay word tokenization. *Journal of Information and Communication Technology*, 15(1):107–131.

Muhammad Fakhrur Razi Abu Bakar, Norisma Idris, and Liyana Shuib. 2019. An enhancement of Malay social media text normalization for lexicon-based sentiment analysis. *23rd International Conference on Asian Language Processing, IALP 2019*, pages 211–215.

Normi Sham Abu Bakar. 2020. The development of an integrated corpus for Malay language. *6th International Conference on Computational Science and Technology, ICCST 2019*, 603:425–433.

Zamri Abu Bakar, Nomaly Kamal Ismail, and Mohd Izani Mohamed Rawi. 2017. Detection of compound word with combination noun and adjective using rule based technique in Malay standard document. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(3-5 Special Issue):129–134.

Zamri Abu Bakar, Normaly Kamal Ismail, and Mohd Izani Mohamed Rawi. 2018a. Identification of noun + verb compound nouns in Malay standard document based on rule based. *3rd IEEE International Conference on Engineering Technologies and Social Sciences, ICETSS 2017*, 2018-January:1–6.

Zamri Abu Bakar, Normaly Kamal Ismail, Mohd Izani Mohamed Rawi, and Nurazzah Abdul Rahman. 2018b. Automatic detection of compound word in Malay standard document using rule based technique. *2017 IEEE Conference on Open Systems, ICOS 2017*, 2018-January:59–64.

Vimala Balakrishnan, Mohammed Kaity, Hajar Abdul Rahim, and Nazari Ismail. 2021. Social media analytics using sentiment and content analyses on the 2018 Malaysia's general election. *Malaysian Journal of Computer Science*, 34(2):171–183.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424. Association for Computational Linguistics.

Thomas Agung Basuki and Benediktus Giovanito Antaputra. 2020a. How similar is similar: A comparison of Bahasa Indonesia and Bahasa Malaysia. *3rd International Conference on Electronics, Communications and Control Engineering, ICECC 2020*, pages 8–12.

Thomas Anung Basuki and Benediktus Giovanito Antaputra. 2020b. How similar is similar: A comparison of Bahasa Indonesia and Bahasa Malaysia. In *Proceedings of the 3rd International Conference on Electronics, Communications and Control Engineering*, ICECC 2020, page 8–12, New York, NY, USA. Association for Computing Machinery.

Shaiful Bakhtiar Bin Rodzman, Mohammad Hanif Rashid, Normaly Kamal Ismail, Nurazzah Abd Rahman, Syed Ahmad Aljunid, and Hayati Abd Rahman. 2019a. Experiment with lexicon based techniques on domain-specific Malay document sentiment analysis. *9th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2019*, pages 330–334.

Shaiful Bakhtiar Bin Rodzman, Nomaly Kamal Ismail, and Nurazzah Abd Rahman. 2018a. A survey on context-aware information retrieval research. *4th International Conference on Computational Science and Technology, ICCST17*, 488:399–409.

Shaiful Bakhtiar Bin Rodzman, Normaly Kamal Ismail, Nurazzah Abd Rahman, and Zulhilmi Mohamed Nor. 2018b. The implementation of fuzzy logic controller for defining the ranking function on Malay text corpus. *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017*, 2018-January:93–98.

Shaiful Bakhtiar Bin Rodzman, Normaly Kamal Ismail, Nurazzah Abd Rahman, Syed Ahmad Aljunid, Hayati Abd Rahman, Zulhilmi Mohamed Nor, Ku Muhammad Naim Ku Khalif, and Ahmad Yunus Mohd Noor. 2019b. Experiment with text summarization as a positive hierarchical fuzzy logic ranking indicator for domain specific retrieval of Malay translated hadith. *9th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2019*, pages 299–304.

Maslida Binti Yusof and Nurul Jamilah Binti Rosly. 2018. Conceptual structure representation of causative verb in Malay language and relation with syntax. *GEMA Online Journal of Language Studies*, 18(4):143–167.

Francis Bond, Hiroki Nomoto, Luis Morgado da Costa, and Arthur Bond. 2020. Linking the TUFS Basic Vocabulary to the Open Multilingual Wordnet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3181–3188. European Language Resources Association.

Prachya Boonkwan, Thepchai Supnithi, Wandee Tosuwan, and Chai Wutiwiwatchai. 2016. The development of an audible Pattani Malay-Thai electronic phrasebook for military purposes. *5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU 2016*, 81:237–242.

Annisa Briliani, Budhi Irawan, and Casi Setianingsih. 2019. Hate speech detection in Indonesian language on Instagram comment section using K-nearest neighbor classification method. *2019 IEEE International Conference on Internet of Things and Intelligence System, IoTaIS 2019*, pages 98–104.

Widodo Budiharto, Vincent Andreas, and Alexander Agung Santoso Gunawan. 2021. A novel model and implementation of humanoid robot with facial expression and natural language processing (NlP). *ICIC Express Letters, Part B: Applications*, 12(3):275–281.

Marvin Jerremy Budiman and Dessi Puji Lestari. 2020. Multi speaker speech synthesis system for Indonesian language. *7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020.*

Sari Dewi Budiwati and Masayoshi Aritsugi. 2019. Multiple pivots in statistical machine translation for low resource languages. *33rd Pacific Asia Conference on Language, Information and Computation, PACLIC 2019*, pages 345–355.

Sakti Putra Perdana Bunga Batara, Budhi Irawan, and Casi Setianingsih. 2019. Hate speech detection in Indonesian language on Instagram comment section using deep neural network classification method. *5th IEEE Asia Pacific Conference on Wireless and Mobile, APWiMob 2019*, pages 143–149.

Ghulam Asrofi Buntoro, Rizal Arifin, Gus Nanang Syaifuddiin, Ali Selamat, O. Krejcar, and H. Fujita. 2021. Implementation of a machine learning algorithm for sentiment analysis of Indonesia's 2019 presidential election. *IIUM Engineering Journal*, 22(1):78–92.

Francesco Burroni, Sireemas Maspong, Pittayawat Pittayaporn, and Pimthip Kochaiyaphum. 2020. A new look at Pattani Malay initial geminates: a statistical and machine learning approach. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 21–29.

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 160–169. Association for Computational Linguistics.

Alson Cahyadi and Masayu Leylia Khodra. 2018. Aspect-based sentiment analysis using convolutional neural network and bidirectional long short-term memory. *5th International Conference on Advanced Informatics: Concepts Theory and Applications, ICAICTA 2018*, pages 124–129.

Denis Eka Cahyani, Langlang Gumilar, and Ajie Pangestu. 2020. Indonesian parsing using Probabilistic Context-Free Grammar (PCFG) and Viterbi-Cocke Younger Kasami (Viterbi-CYK). *3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, pages 56–61.

Denis Eka Cahyani, Ruli Manurung, and Rahmad Mahendra. 2016. Knowledge representation system for copula sentence in Bahasa Indonesia based on Web Ontology Language (OWL). *International Conference on Advanced Computer Science and Information Systems, ICACSIS 2015*, pages 137–142.

Denis Eka Cahyani and Mtchael Juan Vindiyanto. 2019. Indonesian part of speech tagging using hidden markov model - ngram viterbi. *4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019*, pages 353–358.

Elok Cahyaningtyas and Dhany Arifianto. 2018. Development of under-resourced Bahasa Indonesia speech corpus. *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, 2018-February:1097–1101.

Risma Mustika Cahyaningtyas, Retno Kusumaningrum, Sutikno, Suhartono, and Djalal Er Riyanto. 2017. Emotion detection of tweets in Indonesian language using LDA and expression symbol conversion. *1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018-January:253–257.

Zefeng Cai, Nankai Lin, Chuyu Ma, and Shengyi Jiang. 2019. Indonesian automatic text summarization based on a new clustering method in sentence level. *2019 International Conference on Big Data Engineering, BDE 2019*, pages 30–35.

A. Candra, Wella, and A. Wicaksana. 2021. Bidirectional encoder representations from transformers for cyberbullying text detection in Indonesian social media. *International Journal of Innovative Computing, Information and Control*, 17(5):1599–1615.

Henri Chambert-Loir. 2019. The particle pun in modern Indonesian and Malaysian (La particule pun en Indonésien et Malaisien modernes.). *Archipel. Études interdisciplinaires sur le monde insulindien*, (98):177–237.

Reza Chandra, M. Agung Sucipta Iskandar, Lintang Yuniar Banowosari, Adang Suhendra, and Prihandoko Prihandoko. 2019. Building corpus in Bahasa Indonesia for pornographic indicated website content. *5th International Conference on Computing Engineering and Design, ICCED 2019*.

Khalifa Chekima and Rayner Alfred. 2016. An automatic construction of Malay stop words based on aggregation method. *2nd International Conference on Soft Computing in Data Science, SCDS 2016*, 652:180–189.

Khalifa Chekima and Rayner Alfred. 2018. Sentiment analysis of Malay social media text. *4th International Conference on Computational Science and Technology, ICCST17*, 488:205–219.

Khalifa Chekima, Rayner Alfred, and Kim On Chin. 2018. Rule-based model for Malay text sentiment analysis. *4th International Conference on Computational Science and Technology, ICCST17*, 488:172–185.

Feng Chen, Jian Yang, and Lixuan Zhao. 2020. A bilingual speech synthesis system of standard Malay and Indonesian based on HMM-DNN. *2020 International Conference on Asian Language Processing, IALP 2020*, pages 181–186.

Andry Chowanda and Alan Darmasaputra Chowanda. 2017. Recurrent neural network to deep learn conversation in Indonesian. *2nd International Conference on Computer Science and Computational Intelligence, ICCSCI 2017*, 116:579–586.

Andry Chowanda and Alan Darmasaputra Chowanda. 2018. Generative Indonesian conversation model using recurrent neural network with attention mechanism. *3rd International Conference on Computer Science and Computational Intelligence, ICCSCI 2018*, 135:433–440.

Christianto, Julio Christian Young, and Andre Rusli. 2020. Evaluating RNN architectures for handling imbalanced dataset in multi-class text classification in Bahasa Indonesia. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(5):8418–8423.

Chong Chai Chua, Tek Yong Lim, Lay-Ki Soon, Enya Kong Tang, and Bali Ranaivo-Malançon. 2017. Meaning preservation in Example-based Machine Translation with structural semantics. *Expert Systems with Applications*, 78:242–258.

S. Chua and P. N. E. Nohuddin. 2017. Relationship analysis of keyword and chapter in Malay-translated tafseer of Al-Quran. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(2-10):185–189.

Siaw-Fong Chung. 2019. Lagi in standard Malaysian Malay: Its meaning conceptualization. *Concentric: Studies in Linguistics*, 45(1):82–111.

Siaw-Fong Chung and Meng-Hsien Shih. 2019. An annotated news corpus of Malaysian Malay. *Nusa*, pages 7–34.

Paolo Coluzzi. 2017. Language planning for Malay in Malaysia: A case of failure or success? *International Journal of the Sociology of Language*, 2017(244):17–38.

Mohammad Darwich, Shahrul Azman Mohd Noah, and Nazlia Omar. 2017. Minimally-supervised sentiment lexicon induction model: A case study of Malay sentiment analysis. *11th Multi-disciplinary International Workshop on Artificial Intelligence, MIWAI 2017*, 10607 LNAI:225–237.

Robby Darwis, Herry Sujaini, and Rudy Dwi Nyoto. 2019. Peningkatan mesin penerjemah statistik dengan menambah kuantitas korpus monolingual (studi kasus: Bahasa Indonesia-Sunda). *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 7(1):27–32.

Karlina Denistia and R. Harald Baayen. 2019. The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, 29(3):385.

Destiani, Andayani, and Muhammad Rohmadi. 2018a. Vocabulary load on two mainstream Indonesian textbooks for foreign learners: A comparative study. *International Journal of Social Sciences & Educational Studies*, 5(2):137–151.

Andayani Destiani, Andayani Andayani, and Muhammad Rohmadi. 2018b. Perbandingan deiksis pada dua buku ajar: Analisis kontrastif BIPA dan Bahasa Inggris. *Jurnal Pendidikan Bahasa dan Sastra*, 18(2):151–162.

Dyah Ayu Cyntya Dewi, Shaufiah, and Ibnu Asror. 2018. Analysis and implementation of cross lingual short message service spam filtering using graph-based k-nearest neighbor. *International Conference on Data and Information Science 2017, ICoDIS 2017*, 971.

Haru Deliana Dewi, Andika Wijaya, and Rahayu S. Hidayat. 2021. English legalese translation into Indonesian. *Wacana*, 21(3):446–474.

Intan Novita Dewi, Rahmat Nurcahyo, and Farizal. 2020. Word cloud result of mobile payment user review in Indonesia. *7th IEEE International Conference on Industrial Engineering and Applications, ICIEA 2020*, pages 989–992.

Dhammajoti, Julio Christian Young, and Andre Rusli. 2020. A comparison of supervised text classification and resampling techniques for user feedback in Bahasa Indonesia. *5th International Conference on Informatics and Computing, ICIC 2020*.

Fudholi Dhomas Hatta and Juwairi Kiki Purnama. 2021. Classifying medical document in Bahasa Indonesia using semi-supervised learning. *IOP Conference Series. Materials Science and Engineering*, 1077(1).

M. M. Din, N. H. H. Hashim, and M. M. Siraj. 2017. Comparative study on corpus development for Malay investment fraud detection in website. *Journal of Fundamental and Applied Sciences*, 9(6S):828–838.

Arawinda Dinakaramani and Totok Suhardijanto. 2019. Building a web-based application for language resources in Indonesia. *2nd International Conference on Data and Information Science, ICoDIS 2018*, 1192.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2016. Similar southeast Asian languages: Corpus-based case study on Thai-Laotian and Malay-Indonesian. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 149–156. The COLING 2016 Organizing Committee.

Zuraidah Mohd Don and Gerry Knowles. 2020. New tools for old tasks: A new approach to the investigation of Malay. *Journal on Asian Linguistic Anthropology*, 2(3):21–38.

Mohamad Draman, Din Chai Tee, Zainuddin Lambak, Mohd Razman Yahya, Mohd Izwardi Bin Mohd Yusoff, S. H. Ibrahim, Shahril Saidon, N. Abu Haris, and Tien-Ping Tan. 2017. Malay speech corpus of telecommunication call center preparation for ASR. *5th International Conference on Information and Communication Technology, ICoIC7 2017*.

Meisyarah Dwiastuti. 2019. English-Indonesian neural machine translation for spoken language domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 309–314. Association for Computational Linguistics.

Suci Dwijayanti, Muhammad Abid Tami, and Bhakti Yudho Suprapto. 2021. Speech-to-text conversion in Indonesian language using a deep bidirectional long short-term memory algorithm. *International Journal of Advanced Computer Science and Applications*, 12(3):225–230.

Diyan Ermawan Effendi and Muchammadun. 2018. "happiness" in Bahasa Indonesia and its implication to health and community well-being. *Asian EFL Journal*, 20(8):279–291.

Sukmawati Nur Endah, Satriyo Adhy, and Sutikno. 2017. Comparison of feature extraction MFCC and LPC in automatic speech recognition for Indonesian. *Telkomnika (Telecommunication Computing Electronics and Control)*, 15(1):292–298.

Elvira Erizal, Budhi Irawan, and Casi Setianingsih. 2019. Hate speech detection in Indonesian language on Instagram comment section using maximum entropy classification method. *2nd International Conference on Information and Communications Technology, ICOIACT 2019*, pages 533–538.

Muhammad Izzuddin Eshak, Rohiza Ahmad, and Aliza Sarlan. 2018. A preliminary study on hybrid sentiment model for customer purchase intention analysis in social commerce. *2017 IEEE Conference on Big Data and Analytics, ICBDA 2017*, 2018-January:61–66.

Ditari Salsabila Esperanti and Ayu Purwarianti. 2016. Relation extraction using dependency tree kernel for Bahasa Indonesia. *4th IGNITE Conference and 2016 International Conference on Advanced Informatics: Concepts, Theory and Application, ICAICTA 2016*.

Alaba Ayotunde Fadele, Amirrudin Kamsin, Khadher Ahmad, and Rasheed Abubakar Rasheed. 2020. A novel hadith authentication mobile system in Arabic to Malay language translation for android and iOS phones. *International Journal of Information Technology (Singapore)*, 13(4):1683–1692.

Ahmad Fadly. 2018. Pengembangan kamus pemelajar Bahasa Indonesia bagi penutur asing tingkat dasar di Universitas Muhammadiyah Jakarta. *Pena Literasi*, 1(2):74–80.

Fahmi Fahmi, Meganingrum Arista Jiwanggi, and Mirna Adriani. 2020. Speech-emotion detection in an Indonesian movie. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, volume Language Resources and Evaluation Conference (LREC 2020), pages 185–193. European Language Resources Association (ELRA).

Muhammad Fairuzz Fairuzz Hiloh, Mohd Juzaiddin Ab Aziz, and Lailatul Qadri Zakaria. 2018. The effectiveness of bottom up technique with probabilistic approach for a Malay parser. *GEMA Online Journal of Language Studies*, 18(2):124–133.

Edi Faisal, Farza Nurifan, and Riyanarto Sarno. 2018. Word sense disambiguation in Bahasa Indonesia using svm. *3rd International Seminar on Application for Technology of Information and Communication, iSemantic 2018*, pages 239–243.

Ahmad Muammar Fanani and Suyanto Suyanto. 2021. Syllabification model of Indonesian language named-entity using syntactic n-gram. *5th International Conference on Computer Science and Computational Intelligence, ICCSCI 2020*, 179:721–727.

Laina Farsiah, Yi-Shin Chen, and Alim Misbullah. 2020. Multi-classes emotion detection for unbalanced indonesian tweets. *2020 International Conference on Electrical Engineering and Informatics, ICELTICs 2020*, 2020-October.

M. Ali Fauzi. 2018. Random forest approach fo sentiment analysis in Indonesian language. *Indonesian Journal of Electrical Engineering and Computer Science*, 12(1):46–50.

M. Ali Fauzi and Anny Yuniarti. 2018. Ensemble method for Indonesian Twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(1):294–299.

Ridi Ferdiana, William Fajar, Desi Dwi Purwanti, Artmita Sekar Tri Ayu, and Fahim Jatmiko. 2019. Twitter sentiment analysis in under-resourced languages using byte-level recurrent neural model. *International Journal of Advanced Computer Science and Applications*, 10(8):108–112.

Ivan Ferdino and Andre Rusli. 2019. Using naïve bayes classifier for application feedback classification and management in Bahasa Indonesia. *5th International Conference on New Media Studies, CONMEDIA 2019*, pages 217–222.

Mohammad Fikri and Riyanarto Sarno. 2019. A comparative study of sentiment analysis using svm and senti word net. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(3):902–909.

Devi Fitrianah, Dwiki Jatikusumo, and Ida Nurhaida. 2020. D-loc apps: A location detection application based on social media platform in the event of a flood disaster. *2nd Asia Pacific Information Technology Conference, APIT 2020*, pages 41–45.

Novi Sofia Fitriasari, Khalifa Esha Iftitah, and Rizky Rachman Judhie Putra. 2017. Indonesian document retrieval using vector space method. *3rd International Conference on Science in Information Technology, ICSITech 2017*, 2018-January:664–668.

Yingwen Fu, Nankai Lin, Xiaotian Lin, and Shengyi Jiang. 2021. Towards corpus and model: Hierarchical structured-attention-based features for Indonesian named entity recognition. *Journal of Intelligent & Fuzzy Systems*, 41(1):1–12.

Shamsan Gaber, Mohd Zakree Ahmad Nazri, Nazlia Omar, and Salwani Abdullah. 2020. Part-of-speech (pos) tagger for Malay language using naÏve bayes and k-nearest neighbor model. *Journal of Critical Reviews*, 7(16):248–257.

Garmastewira Garmastewira and Masayu Leylia Khodra. 2019. Summarizing Indonesian news articles using graph convolutional network. *Journal of Information and Communication Technology*, 18(3):345–365.

Beat Gfeller, Vlad Schogol, and Keith Hall. 2016. Cross-lingual projection for class-based language models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 83–88.

Hishshah Ghassani and Tricya Esterina Widagdo. 2018. Access to relational databases using interrogative sentences in Indonesian language. *5th International Conference on Data and Software Engineering, ICoDSE 2018*.

Yohanes Gultom and Wahyu Catur Wibowo. 2018. Automatic open domain information extraction from Indonesian text. *2017 International Workshop on Big Data and Information Security, WBIS 2017*, 2018-January:23–30.

Gunarso Gunarso, Hammam Riza, Elvira Nurfadhilah, M. Teduh Uliniansyah, Agung Santosa, and Lyla R. Aini. 2016. An overview of BPPT's Indonesian language resources. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 73–77.

D. Gunawan and A. Amalia. 2017. The design of lexical database for Indonesian language. *IOP Conference Series. Materials Science and Engineering*, 180(1).

D. Gunawan, A. Amalia, M. S. Lydia, and M. I. Muthaqin. 2018a. The observation of Bahasa Indonesia official computer terms implementation in scientific publication. *Journal of Physics: Conference Series*, 979(1).

D. Gunawan, A. Amalia, and O. N. Maringga. 2019a. Building the application to identify incorrect capital letters writing in Bahasa Indonesia. *Journal of Physics: Conference Series*, 1235(1).

D. Gunawan, A. Pasaribu, R. F. Rahmat, and R. Budiarto. 2017a. Automatic text summarization for Indonesian language using textteaser. *IOP Conference Series. Materials Science and Engineering*, 190(1).

Dani Gunawan and Amalia Amalia. 2018. Review of the recent research on automatic text summarization in Bahasa Indonesia. *3rd International Conference on Informatics and Computing, ICIC 2018*.

Dani Gunawan, Amalia Amalia, and Indra Charisma. 2017b. Automatic extraction of multiword expression candidates for Indonesian language. *6th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2016*, pages 304–309.

Dani Gunawan, Siti Hazizah Harahap, and Romi Fadillah Fadillah Rahmat. 2019b. Multi-document summarization by using textrank and maximal marginal relevance for text in Bahasa Indonesia. *10th International Conference on ICT for Smart Society, ICISS 2019*.

Dani Gunawan, Syaiful Anwar Husen Lubis, Romi Fadillah Rahmat, and Ainul Hizriadi. 2019c. Building the pornography corpus for Bahasa Indonesia based on TRUST+™ positif database. *10th International Conference on ICT for Smart Society, ICISS 2019*.

Dani Gunawan, Fanindia Purnamasari, Ranti Ramadhiana, and Romi Fadillah Rahmat. 2020. Keyword extraction from scientific articles in Bahasa Indonesia using textrank algorithm. *4th International Conference on Electrical, Telecommunication and Computer Engineering, ELTICOM 2020*, pages 260–264.

Dani Gunawan, Hardiani Putri Siregar, and Opim Salim Sitompul. 2019d. Identifying sentence structure in Bahasa Indonesia by using pos tag and lalr parser. *5th International Conference on Computing Engineering and Design, ICCED 2019*.

Deri Gunawan, Rendra Mahardika, Feri Ranja, Sarah Purnamawati, and Ivan Jaya. 2019e. The identification of pornographic sentences in Bahasa Indonesia. *5th Information Systems International Conference, ISICO 2019*, 161:601–606.

Reza Gunawan, Ichan Taufik, Edi Mulyana, Opik Taupik Kurahman, Muhammad Ali Ramdhani, and Mahmud Mahmud. 2019f. Chatbot application on internet of things (iot) to support smart urban agriculture. *5th International Conference on Wireless and Telematics, ICWT 2019*.

Teddy Surya Gunawan, Rashida Husain, and Mira Kartiwi. 2018b. Development of language identification system using mfcc and vector quantization. *4th IEEE International Conference on Smart Instrumentation, Measurement and Applications, ICSIMA 2017*, 2017-November:1–4.

William Gunawan, Derwin Suhartono, Fredy Purnomo, and Andrew Ongko. 2018c. Named-entity recognition for Indonesian language using bidirectional lstm-cnns. *3rd International Conference on Computer Science and Computational Intelligence, ICCSCI 2018*, 135:425–432.

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43.

Abid Nurul Hakim, Rahmad Mahendra, Mima Adriani, and Adrianus Saga Ekakristi. 2018. Corpus development for Indonesian consumer-health question answering system. *9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018-January:222–227.

Harnisa Azrin Hakimi and Nurazzah Abd Rahman. 2021. Developing the covid-19 Malay corpus using wordpress content management system (cms). In *5th International Conference on Information Retrieval and Knowledge Management, CAMP 2021*, pages 75–83. Institute of Electrical and Electronics Engineers Inc.

Christian Halim, Alfan Farizki. Wicaksono, and Mirna Adriani. 2018. Extracting disease-symptom relationships from health question and answer forum. *21st International Conference on Asian Language Processing, IALP 2017*, 2018-January:87–90.

Mohd Pouzi Hamzah and Syarifah Fatem Na'imah Binti Syed Kamaruddin. 2021. Open text ontology mining to improve retrievals of information. *International Journal of Advanced Computer Science and Applications*, 12(7):504–511.

Raseeda Hamzah, Nursuriati Jamil, Khyrina Airin Fariza Abu Samah, Nur Nabilah Abu Mangshor, Nurbaity Sabri, and Rosniza Roslan. 2017. Comparing statistical classifiers for emotion classification. *7th IEEE International Conference on System Engineering and Technology, ICSET 2017*, pages 183–188.

Novita Hanafiah, Alexander Kevin, Charles Sutanto, Fiona, Yulyani Arifin, and Jaka Hartanto. 2017. Text normalization algorithm on Twitter in complaint category. *2nd International Conference on Computer Science and Computational Intelligence, ICCSCI 2017*, 116:20–26.

Dini Handayani, Normi Sham Awang Abu Bakar, Hamwira Yaacob, and Mustafa Ali Abuzaraida. 2018. Sentiment analysis for Malay language: systematic literature review. *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, pages 305–310.

Dellon Handrata, Christian Nathaniel Purwanto, Fransisca Haryanti Chandra, Joan Santoso, and Gunawan. 2019. Part of speech tagging for Indonesian language using bidirectional long short-term memory. *1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, pages 85–88.

Rafizah Mohd Hanifa, Khalid Isa, Shamsul Mohamad, Shaharil Moh Shah, Shelena Soosay Nathan, Rosni Ramle, and Mazniha Berahim. 2019. Voiced and unvoiced separation in Malay speech using zero crossing rate and energy. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(2):775–780.

Haslizatul Mohamed Hanum and Zainab Abu Bakar. 2016a. Evaluation of energy and duration on Malay phrase breaks. *9th Asia International Conference on Mathematical Modelling and Computer Simulation - Asia Modelling Symposium, AMS 2015*, pages 101–104.

Haslizatul Mohamed Hanum and Zainab Abu Bakar. 2016b. Sentence segmentation and phrase strength estimation in Malay continuous speech. *8th Speech Prosody 2016*, 2016-January:1163–1166.

Haslizatul Mohamed Hanum, Syazwani Nasaruddin, and Zainab Abu Bakar. 2017. Prosodic breaks on Malay speech corpus: Evaluation of pitch, intensity and duration. *3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016*, pages 43–47.

Haslizatul Mohamed Hanum, Nur Farhana Rasip, and Zainab Abu Bakar. 2019. Multi-word similarity and retrieval model for a refined retrieval of quranic sentences. *6th International Conference on Advances in Visual Informatics, IVIC 2019*, 11870 LNCS:380–389.

Mukhtar Haris, Moch Arif Bijaksana, and Totok Suhardijanto. 2019. Warning and suggestion system on syntax tree maker application. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 113–117.

Hanny Haryanto and Aripin. 2019. A finite state machine model to determine syllables of Indonesian text. *1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, pages 238–241.

Uswatun Hasanah, Tri Astuti, Rizki Wahyudi, Zanuar Rifai, and Rilas Agung Pambudi. 2018. An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian. *3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, pages 230–234.

Siti Zubaidah Mohd Hashim and Hajar Abdul Rahim. 2016. Defying the global: The cultural connotations of "Islam" in Malaysia. *Kemanusiaan*, 23(Supp. 2):81–98.

Ramos Janoah Hasudungan and Ayu Purwarianti. 2019. Relation detection for Indonesian language using deep neural network - support vector machine. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 290–296.

Mohamad Hazim, Nor Badrul Anuar, Mohd Faizal Ab Razak, and Nor Aniza Abdullah. 2018. Detecting opinion spams through supervised boosting approach. *PLoS ONE*, 13(6):e0198884.

Alex Henry and Debbie G. E. Ho. 2016. Code-switching in bruneian online retail transactions. *World Englishes*, 35(4):554–570.

Herlawati, Rahmadya Trias Handayanto, Didik Setiyadi, and Endang Retnoningsih. 2019. Corpus usage for sentiment analysis of a hashtag Twitter. *4th International Conference on Informatics and Computing, ICIC 2019*.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576.

Risanuri Hidayat, Priyatmadi, and Welly Ikawijaya. 2016. Wavelet based feature extraction for the vowel sound. *2nd International Conference on Information Technology Systems and Innovation, ICITSI 2015*.

A. F. Hidayatullah and M. R. Ma'arif. 2017. Preprocessing tasks in Indonesian Twitter messages. *Journal of Physics: Conference Series*, 801(1).

Ahmad Fathan Hidayatullah, Wisnu Kurniawan, and Chanifah Indah Ratnasari. 2019. Topic modeling on Indonesian online shop chat. *3rd International Conference on Natural Language Processing and Information Retrieval, NLPIR 2019*, pages 121–126.

Ahmad Fathan Hidayatullah, Elang Cergas Pembrani, Wisnu Kurniawan, Gilang Akbar, and Ridwan Pranata. 2018. Twitter topic modeling on football news. *3rd International Conference on Computer and Communication Systems, ICCCS 2018*, pages 94–98.

Ahmad Fathan Hidayatullah, Chanifah Indah Ratnasari, and Satrio Wisnugroho. 2016. Analysis of stemming influence on Indonesian tweet classification. *Telkomnika (Telecommunication Computing Electronics and Control)*, 14(2):665–673.

Satria Nur Hidayatullah and Suyanto. 2019. Developing an adaptive language model for Bahasa Indonesia. *International Journal of Advanced Computer Science and Applications*, 10(1):488–492.

Mohd Hanafi Ahmad Hijazi, Lyndia Libin, Rayner Alfred, and Frans Coenen. 2017. Bias aware lexicon-based sentiment analysis of Malay dialect on social media data: A study on the Sabah language. *2nd International Conference on Science in Information Technology, ICSITech 2016*, pages 356–361.

Awaliyatul Hikmah, Sumarni Adi, and Mulia Sulistiyono. 2020. The best parameter tuning on rnn layers for Indonesian text classification. *3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, pages 94–99.

Xavier Hinaut and Johannes Twiefel. 2020. Teach your robot your language! Trainable neural parser for modeling human sentence processing: Examples for 15 languages. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):179–188.

Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Devin Hoesen, Dessi Puji Lestari, and Dwi Hendratmo Widyantoro. 2018. Shared-hidden-layer deep neural network for under-resourced language the content. *Telkomnika (Telecommunication Computing Electronics and Control)*, 16(3):1226–1238.

Devin Hoesen, Fanda Yuliana Putri, and Dessi Puji Lestari. 2019. Automatic pronunciation generator for Indonesian speech recognition system based on sequence-to-sequence model. *22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2019*.

Tom G. Hoogervorst. 2018. Utterance-final particles in Klang Valley Malay. *Wacana*, 19(2):291–326.

Ahmad Hany Hossny and Lewis Mitchell. 2019. Event detection in Twitter: A keyword volume approach. *18th IEEE International Conference on Data Mining Workshops, ICDMW 2018*, 2018-November:1200–1208.

Tan Kim Hua, Hamdi Khalis, Nur Ehsan Mohd-Said, and Ong Song Howe. 2021. The polarity of war metaphors in sports news: A corpus-informed analysis. *GEMA Online Journal of Language Studies*, 21(2):238–252.

Tan Kim Hua, Shahidatul Maslina Mat So'od, and Bahiyah Abdul Hamid. 2019. Communicating insults in cyberbullying. *SEARCH (Malaysia)*, 11(3):91–109.

Khodijah Hulliyah, Husni Teja Sukmana, Normi Sham Abu Bakar, and Amelia Ritahani Ismail. 2019. Indonesian affective word resources construction in valence and arousal dimension for sentiment analysis. *6th International Conference on Cyber and IT Service Management, CITSM 2018*.

Khodijah Hulliyah, Abdul Wahab, Norhaslinda Kamaruddin, Sevki Erdogan, and Yusuf Durachman. 2017. Analysis of Indonesian sentiment text based on affective space model (ASM) using electroencephalogram (EEG) signals. *1st International Conference on Informatics and Computing, ICIC 2016*, pages 325–328.

Mohd Zabidin Husin, Saidah Saad, and Shahrul Azman Mohd Noah. 2018. Syntactic rule-based approach for extracting concepts from quranic translation text. *6th International Conference on Electrical Engineering and Informatics, ICEEI 2017*, 2017-November:1–6.

Husni, Ika Oktavia Suzanti, Yoga Dwitya Pramudita, Putro Sigit Susanto, and Lukman Heryawan. 2020. Web service for search engine Bahasa Indonesia (sebi). *Journal of Physics: Conference Series*, 1569(2).

Amalia Asti Hutami, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Paraphrase construction of Al Quran in Indonesian language translation. *7th International Conference on Information and Communication Technology, ICoICT 2019*.

Jacqueline Ibrahim and Dessi Puji Lestari. 2018. Classification and clustering to identify spoken dialects in Indonesian. *4th International Conference on Data and Software Engineering, ICoDSE 2017*, 2018-January:1–6.

Noor Jamaliah Ibrahim, Mohd Yamani Idna Idris, Mohd Yakub Zulkifli Mohd Yusoff, Noor Naemah Abdul Rahman, and Mawil Izzi Dien. 2019. Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. *Malaysian Journal of Computer Science*, 2019(Special Issue 3):46–72.

Muhammad Okky Ibrohim and Indra Budi. 2019a. Multi-label hate speech and abusive language detection in Indonesian Twitter. Proceedings of the Third Workshop on Abusive Language Online, pages 46–57. Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019b. Translated vs non-translated method for multilingual hate speech identification in Twitter. *International Journal on Advanced Science, Engineering and Information Technology*, 9(4):1116–1123.

Muhammad Okky Ibrohim, Muhammad Akbar Setiadi, and Indra Budi. 2019. Identification of hate speech and abusive language on Indonesian Twitter using the word2vec, part of speech and emoji features. *2019 International Conference on Advanced Information Science and System, AISS 2019*.

Nur Oktavin Idris, Widyawan, and Teguh Bharata Adji. 2019. Classification of radicalism content from Twitter written in Indonesian language using long short term memory. *3rd International Conference on Informatics and Computational Sciences, ICICOS 2019*.

M. Ikhwan Syafiq, M. Shukor Talib, Naomie Salim, Habibollah Haron, and Razana Alwee. 2019. A concise review of named entity recognition system: Methods and features. *International Conference on Green Engineering Technology and Applied Computing 2019, IConGETech2 019 and International Conference on Applied Computing 2019, ICAC 2019*, 551.

Helmi Imaduddin, Widyawan, and Silmi Fauziati. 2019. Word embedding comparison for Indonesian language sentiment analysis. *1st International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pages 426–430.

Imamah, Husni, Eka Malasari Rachman, Ika Oktavia Suzanti, and Fifin Ayu Mufarroha. 2020. Text mining and support vector machine for sentiment analysis of tourist reviews in bangkalan regency. *Journal of Physics: Conference Series*, 1477(2).

Zul Indra, Jafreezal Jaafar, Norshuhani Zamin, and Zainab Abu Bakar. 2016. A language identifier for Indonesian and Malay text document. *2015 International Symposium on Mathematical Sciences and Computing Research, iSMSC 2015*, pages 127–131.

Budi Irmawati, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Exploiting syntactic similarities for preposition error corrections on Indonesian sentences written by second language learner. *5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU 2016*, 81:214–220.

Budi Irmawati, Hiroyuki Shindo, and Yuji Matsumoto. 2017a. A dependency annotation scheme to extract syntactic features in Indonesian sentences. *International Journal of Technology*, 8(5):957–967.

Budi Irmawati, Hiroyuki Shindo, and Yuji Matsumoto. 2017b. Generating artificial error data for Indonesian preposition error corrections. *International Journal of Technology*, 8(3):549–558.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273.

Asiah Ismail, Anida Sarudin, Zulkifli Osman, and Husna Faredza Mohamed Redzwan. 2021. The process of forming a more complex idiomatic meaning using a principle of integration metaphors. *GEMA Online Journal of Language Studies*, 21(2):86–110.

B. H. Iswanto and V. Poerwoto. 2018. Sentiment analysis on Bahasa Indonesia tweets using unibigram models and machine learning techniques. *IOP Conference Series. Materials Science and Engineering*, 434(1).

Jafreezal Jaafar, Zul Indra, and Nurshuhaini Zamin. 2016. A category classification algorithm for Indonesian and Malay news documents. *Jurnal Teknologi*, 78(8-2):121–132.

Aris Tri Jaka Harjanta and Bambang Agus Herlambang. 2020. Extraction sentiment analysis using naive bayes algorithm and reducing noise word applied in Indonesian language. *7th International Conference on DV-Xa Method: The Advances-Related Experiments and Theories on Material Science, ICDM 2019*, 835.

Norezmi Jamal, N Fuad, and M. N. A. H. Sha'abani. 2020. A hybrid approach for single channel speech enhancement using deep neural network and harmonic regeneration noise reduction. *International Journal of Advanced Computer Science and Applications*, 11(10):243–248.

Muhammad Nabil Fikri Jamaluddin, Siti Zaleha Zainal Abidin, and Nasiroh Omar. 2017. Classification and quantification of user's emotion on Malay language in social network sites using latent semantic analysis. *2016 IEEE Conference on Open Systems, ICOS 2016*, pages 65–70.

Muhammad Ihsan Jambak, Fathey Mohammed, Novita Hidayati, Rusdi Efendi, and Rifkie Primartha. 2019. The impacts of singular value decomposition algorithm toward Indonesian language text documents clustering. *3rd International Conference of Reliable Information and Communication Technology, IRICT 2018*, 843:173–183.

Muhammad Ihsan Jambak and Putri Sanggabuana Setiawan. 2018. The development of Bahasa Indonesia corpora for machine learning model in combating cyber bullying: A case study of the Indonesian 2017 capital city governor election. *Journal of Theoretical*

and Applied Information Technology, 96(7):1971–1988.

Nursuriati Jamil, Farihah Apandi, and Raseeda Hamzah. 2017. Influences of age in emotion recognition of spontaneous speech: A case of an under-resourced language. *9th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2017*.

Nurul Syafidah Jamil, Siti Sakira Kamaruddin, and Farzana Kabir Ahmad. 2019. Social tension and crime related events detection method on Twitter. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(6):2821–2824.

Laura A. Janda. 2022. Managing Data and Statistical Code According to the FAIR Principles. In *The Open Handbook of Linguistic Data Management*. The MIT Press.

S. Jiang, S. Li, S. Fu, and N. Lin. 2020. An overview of natural language processing for Indonesian and Malay. *Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence*, 33(6):530–541.

Meganingrum Arista Jiwanggi and Mirna Adriani. 2016. Topic summarization of microblog document in Bahasa Indonesia using the phrase reinforcement algorithm. *5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU 2016*, 81:229–236.

Andreas Jodhinata and Lusia Hartanti. 2016. Naïve bayes implementation into Bahasa Indonesia stemmer for content based webpage classification. *International Journal of Applied Business and Economic Research*, 14(11):8211–8223.

Mohammed Kaity and Vimala Balakrishnan. 2020. An integrated semi-automated framework for domain-based polarity words extraction from an unannotated non-English corpus. *Journal of Supercomputing*, 76(12):9772–9799.

Constantijn Kaland and Stefan Baumann. 2020. Demarcating and highlighting in Papuan Malay phrase prosody. *Journal of the Acoustical Society of America*, 147(4):2974–2988.

Constantijn Kaland and Nikolaus P. Himmelmann. 2020. Repetition reduction revisited: The prosody of repeated words in Papuan Malay. *Language and Speech*, 63(1):31–55.

Constantijn Kaland, Nikolaus P. Himmelmann, and Angela Kluge. 2019. Stress predictors in a Papuan Malay random forest. In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 2871–2875.

Constantijn Kaland, Angela Kluge, and Vincent J. Van Heuven. 2021. Lexical analyses of the function and phonology of Papuan Malay word stress. *Phonetica*, 78(2):141–168.

Syarifah Fatem Na'imah Binti Syed Kamaruddin, Fatihah Mohd, Mohd Pouzi Hamzah, Fadilah Harun, Noor Raihani Zainol, and Nurul Izyan Mat Daud. 2021. Information retrieval for Malay text: A decade review of research (2008-2019). In *5th International Conference on Information Retrieval and Knowledge Management, CAMP 2021*, pages 2–7. Institute of Electrical and Electronics Engineers Inc.

Rathimala Kannan, Ki Soon Lay, and Menagaeswary Govindasamy. 2019. Review on the role of social media for dengue prevention and monitoring. *Applied Mechanics and Materials*, 892:228–233.

Yeni Karlina, Amin Rahman, and Raqib Chowdhury. 2020. Designing phonetic alphabet for Bahasa Indonesia (PABI) for the teaching of intelligible English pronunciation in indonesia. *Indonesian Journal of Applied Linguistics*, 9(3):724–732.

Junaini Kasdan, Rusmadi Baharuddin, and Anis Shahira Shamsuri. 2020. Covid-19 dalam korpus peristilahan Bahasa Melayu: Analisis sosioterminologi (Covid-19 in the corpus of Malay terminology: A socio-terminological analysis). *GEMA Online Journal of Language Studies*, 20(3):221–241.

Junaini Kasdan, Harshita Aini Haroon, Nor Suhaila Che Pa, and Zuhairah Idrus. 2017. Gandaan separa dalam terminologi Bahasa Melayu: Analisis sosioterminologi (Partial reduplication in Malay terminology: A socio-terminological analysis). *GEMA Online Journal of Language Studies*, 17(1):183–202.

Emaliana Kasmuri and Halizah Basiron. 2019. Building a Malay-English code-switching subjectivity corpus for sentiment analysis. *International Journal of Advances in Soft Computing and its Applications*, 11(1):112–130.

Emaliana Kasmuri and Halizah Basiron. 2020. Segregation of code-switching sentences using rule-based technique. *International Journal of Advances in Soft Computing and its Applications*, 12(1):49–64.

Mohamad Nizam Kassim, Shaifal Hisham Mat Jali, Mohd Aizaini Maarof, and Anazida Zainal. 2019. Towards stemming error reduction for Malay texts. *5th International Conference on Computational Science and Technology, ICCST 2018*, 481:13–23.

Mohamad Nizam Kassim, Shaifal Hisham Mat Jali, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2020a. Design consideration of Malay text stemmer using structured approach. In *3rd International Conference on Smart Trends for Information Technology and Computer Communications, SmartCom 2019*, volume 165, pages 421–432. Springer.

Mohamad Nizam Kassim, Shaifal Hisham Mat Jali, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2020b. Enhanced text stemmer with noisy text normalization for Malay texts. In *3rd International Conference on Smart Trends for*

Information Technology and Computer Communications, SmartCom 2019, volume 165, pages 433–444. Springer.

Mohamad Nizam Kassim, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2016a. Enhanced rules application order to stem affixation, reduplication and compounding words in Malay texts. *14th International Workshop on Knowledge Management and Acquisition for Intelligent Systems, PKAW2016*, 9806 LNCS:71–85.

Mohamad Nizam Kassim, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2016b. Malay word stemmer to stem standard and slang word patterns on social media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9714 LNCS:391–400.

Mohamad Nizam Kassim, Mohd Aizaini Maarof, Anazida Zainal, and Amirudin Abdul Wahab. 2016c. Word stemming challenges in Malay texts: A literature review. *4th International Conference on Information and Communication Technology, ICoICT 2016*.

Wandeep Kaur and Vimala Balakrishnan. 2016. Bilingual sentiment detection - investigating impact of tweet translation. *7th International Conference on Applications of Digital Information and Web Technologies, ICADIWT 2016*, 282:105–111.

Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. 2020. Towards a standardized dataset on Indonesian named entity recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 64–71. Association for Computational Linguistics.

Nur Firza Shafiqa Khalid and Normaliza Abd Rahim. 2021. Pola penggunaan Bahasa Melayu dalam Twitter mantan Perdana Menteri ke-enam, Dato' Seri Najib Razak (Patterns of Malay language usage on Twitter of former sixth Prime Minister, Dato' Seri Najib Razak). *Jurnal Komunikasi: Malaysian Journal of Communication*, 37(2):195–209.

Yen-Min Jasmina Khaw, Tien-Ping Tan, and Bali Ranaivo-Malançon. 2017. Automatic phoneme identification for Malay dialects. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(2-9):85–94.

Lau Su Kia and Awab Su'Ad. 2019. A study of education-related Chinese words used in Malaysia-based computer corpus. *Kajian Malaysia*, 37(1):83–107.

Xuan Kong and Jian Yang. 2018. Indonesian corpus constructing and text processing for speech synthesis. In *2018 International Conference on Asian Language Processing (IALP)*, pages 193–196. IEEE.

Fajri Koto. 2016. A publicly available Indonesian corpora for automatic abstractive and extractive chat summarization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 801–805.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian nlp. Proceedings of the 28th International Conference on Computational Linguistics, page 7. International Committee on Computational Linguistics.

Dinar Ajeng Kristiyanti, Akhmad Hairul Umam, Mochamad Wahyudi, Ruhul Amin, and Linda Marlinda. 2019. Comparison of svm naïve bayes algorithm for sentiment analysis toward West Java governor candidate period 2018-2023 based on public opinion on Twitter. *6th International Conference on Cyber and IT Service Management, CITSM 2018*.

Deffri Kun Indarta and Ade Romadhony. 2021. Aspect and opinion extraction of Indonesian lipsticks product reviews using conditional random field (crf). *13th International Conference Knowledge and Smart Technology, KST 2021*, pages 113–117.

Rafly Indra Kurnia and Abba Suganda Girsang. 2021. Classification of user comment using word2vec and deep learning. *International Journal of Emerging Technology and Advanced Engineering*, 11(5):1–8.

Lilis Kurniasari and Arief Setyanto. 2020a. Sentiment analysis using recurrent neural network-lstm in Bahasa Indonesia. *Journal of Engineering Science and Technology*, 15(5):3242–3256.

Lilis Kurniasari and Arif Setyanto. 2020b. Sentiment analysis using recurrent neural network. *Journal of Physics: Conference Series*, 1471(1).

Farhan Wahyu Kurniawan and Warih Maharani. 2020. Indonesian Twitter sentiment analysis using word2vec. *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*.

Kemal Kurniawan and Samuel Louvan. 2018. Empirical evaluation of character-based model on neural named-entity recognition in Indonesian conversational texts.

In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 85–92. Association for Computational Linguistics.

Rahmad Kurniawan, Fitra Lestari, Abdul Somad Batubara, Mohd Zakree Ahmad Nazri, Khairunnas Rajab, and Rinaldi Munir. 2021. Indonesian lexicon-based sentiment analysis of online religious lectures review. In *2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021*. Institute of Electrical and Electronics Engineers Inc.

Selvia Ferdiana Kusuma, Daniel Siahaan, and Umi Laili Yuhana. 2016. Automatic Indonesia's questions classification based on bloom's taxonomy using natural language processing a preliminary study. *2nd International Conference on Information Technology Systems and Innovation, ICITSI 2015*.

Renny Pradina Kusumawardani and Siti Oryza Khairunnisa. 2019. Author-topic modelling for reviewer assignment of scientific papers in Bahasa Indonesia. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 351–356.

Renny Pradina Kusumawardani and Muhammad Wildan Maulidani. 2020. Aspect-level sentiment analysis for social media data in the political domain using hierarchical attention and position embeddings. *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*.

Renny Pradina Kusumawardani, Stezar Priansya, and Faizal Johan Atletiko. 2018. Context-sensitive normalization of social media text in Bahasa Indonesia based on neural word embeddings. *3rd International Neural Network Society Conference on Big Data and Deep Learning, INNS BDDL 2018*, 144:105–117.

Deny A. Kwary. 2019. A corpus platform of Indonesian academic language. *SoftwareX*, 9:102–106.

Teguh Puji Laksono, Ahmad Fathan Hidayatullah, and Chanifah Indah Ratnasari. 2019. Speech to text of patient complaints for Bahasa Indonesia. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 79–84.

Tang Sui Lan and Rajasvaran Logeswaran. 2020. Challenges and development in Malay natural language processing. *Journal of Critical Reviews*, 7(3):61–65.

Anisa Larassati, Nina Setyaningsih, Raden Arief Nugroho, Valentina Widya Suryaningtyas, Setyo Prasiyanto Cahyono, and Stephani Diah Pamelasari. 2019. Google vs. Instagram machine translation: Multilingual application program interface errors in translating procedure text genre. *2019 International Seminar on Application for Technology of Information and Communication, iSemantic 2019*, pages 554–558.

Jeremia Jason Lasiman and Dessi Puji Lestari. 2019. Speech emotion recognition for Indonesian language using long short-term memory. *6th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2018*, pages 40–43.

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131.

Joanna Chiew Ling Lee, Phoey Lee Teh, Sian Lun Lau, and Irina Pak. 2019. Compilation of Malay criminological terms from online news. *Indonesian Journal of Electrical Engineering and Computer Science*, 15(1):355–364.

Sungyoon Lee, Li-Jen Kuo, Zhihong Xu, and Xueyan Hu. 2020. The effects of technology-integrated classroom instruction on k-12 english language learners' literacy development: a meta-analysis. *Computer Assisted Language Learning*, 0(0):1–32.

Rezka Aufar Leonandya, Bayu Distiawan, and Nursidik Heru Praptono. 2016. A semi-supervised algorithm for Indonesian named entity recognition. *3rd International Symposium on Computational and Business Intelligence, ISCBI 2015*, pages 45–50.

Boon Pang Lim, Faith Wong, Yuyao Li, and Jia Wei Bay. 2016. Transfer learning with bottleneck feature networks for whispered speech recognition. *17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, 08-12-September-2016:1578–1582.

Hui Ting Lim, Sharin Hazlin Huspi, and Roliana Ibrahim. 2021. A conceptual framework for Malay-English mixed-language question answering system. In *2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021*. Institute of Electrical and Electronics Engineers Inc.

Shin Huei Lim and Terry Halpin. 2016. Automated verbalization of orm models in Malay and Mandarin. *International Journal of Information System Modeling and Design*, 7(4):1–16.

Nankai Lin, Sihui Fu, Jiawen Huang, and Shengyi Jiang. 2019a. Exploring letter's differences between partial Indonesian branch language and English. *23rd International Conference on Asian Language Processing, IALP 2019*, pages 84–89.

Nankai Lin, Sihui Fu, Shengyi Jiang, Chen Chen, Lixian Xiao, and Gangqin Zhu. 2019b. Learning Indonesian frequently used vocabulary from large-scale news. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 234–239.

Nankai Lin, Sihui Fu, Shengyi Jiang, Gangqin Zhu, and Yanni Hou. 2019c. Exploring lexical differences between Indonesian and Malay. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 178–183.

Chen Liu, Anderson De Andrade, and Muhammad Osama. 2019a. Exploring multilingual syntactic sentence representations. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 153–159.

Wuying Liu and Lin Wang. 2019. Malay-corpus-enhanced Indonesian-Chinese neural machine translation. *10th International Symposium on Intelligence Computation and Applications, ISICA 2018*, 986:239–248.

Wuying Liu and Lin Wang. 2020. Transfer building of multiword expression resource from Indonesian to Malay. *2020 International Conference on Asian Language Processing, IALP 2020*, pages 299–304.

Wuying Liu, Lin Xiao, Shengyi Jiang, and Lin Wang. 2019b. Language resource extension for Indonesian-Chinese machine translation. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 221–225.

Chek Kim Loi and Jason Miin-Hwa Lim. 2019. Hedging in the discussion sections of English and Malay educational research articles. *GEMA Online Journal of Language Studies*, 19(1):36–61.

Yu-Zane Low, Lay-Ki Soon, and Shageenderan Sapai. 2020. A neural machine translation approach for translating Malay parliament hansard to English text. *2020 International Conference on Asian Language Processing, IALP 2020*, pages 316–320.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250. Association for Computational Linguistics.

Made Raharja Surya Mahadi, Anditya Arifianto, and Kurniawan Nur Ramadhani. 2020. Adaptive attention generation for Indonesian image captioning. *8th International Conference on Information and Communication Technology, ICoICT 2020*.

Nurul Husna Mahadzir, Mohd Faizal Omar, and Mohd Nasrun Mohd Nawi. 2018. Semantic similarity measures for Malay-English ambiguous words. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(1-11):109–112.

Rahmad Mahendra, Abid Nurul Hakim, and Mirna Adriani. 2018a. Towards question identification from online healthcare consultation forum post in bahasa. *21st International Conference on Asian Language Processing, IALP 2017*, 2018-January:399–402.

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and Mirna Adriani. 2018b. Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task. In *Proceedings of the 9th Global Wordnet Conference*, pages 245–250.

Syiti Liviani Mahfiz and Ade Romadhony. 2020. Aspect-based opinion mining on beauty product reviews. *3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, pages 488–493.

Nurul Hashimah Ahamed Hassain Malim, Saravanan Sagadevan, and Nurul Izzati Ridzuwan. 2019. Criminality recognition using machine learning on Malay language tweets. *Pertanika Journal of Science and Technology*, 27(4):1803–1820.

Gamaria Mandar and Gunawan Gunawan. 2017. Peringkasan dokumen berita Bahasa Indonesia menggunakan metode cross latent semantic analysis. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 3(2):94–104.

Lindung Parningotan Manik, Hani Febri Mustika, Zaenal Akbar, Yulia Aris Kartika, Dadan Ridwan Saleh, Foni Agus Setiawan, and Ika Atman Satya. 2020. Aspect-based sentiment analysis on candidate character traits in Indonesian presidential election. *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications, ICRAMET 2020*, pages 224–228.

Lindung Parningotan Manik, Arida Ferti Syafiandini, Hani Febri Mustika, Achmad Fatchuttamam Abka, and Yan Rianto. 2019. Evaluating the morphological and capitalization features for word embedding-based pos tagger in Bahasa Indonesia. *6th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2018*, pages 49–53.

N. A. W. Mansor and Nor Hashimah Jalaluddin. 2016. The implicit meaning in Malay figurative language: Synergising communication, cognition and semantics. *Jurnal Komunikasi: Malaysian Journal of Communication*, 32(1):189–206.

Teddy Mantoro, Media Anugerah Ayu, and Rahmadya Trias Handayanto. 2020. Machine learning approach for sentiment analysis in crime information retrieval. *3rd International Conference on Computer and Informatics Engineering, IC2IE 2020*, pages 96–100.

Ngalim Markhamah Abdul, Muhammad Muinudinillah Basri, and Atiqa Sabardila. 2017. Comparison of personal pronoun between Arabic and its Indonesian translation of Koran. *International Journal of Applied Linguistics & English Literature*, 6(5):238–254.

Marlyn Maseri and Mazlina Mamat. 2019. Malay language speech recognition for preschool children using Hidden Markov Model (HMM) system training. *Computational Science and Technology*, pages 205–214.

Ruhaila Maskat, Muhammad Faizzuddin Zainal, Nurissammimayantie Ismail, Norizah Ardi, Amirah Ahmad, and Norizah Daud. 2020. Automatic labelling of Malay cyberbullying Twitter corpus using combinations of sentiment, emotion and toxicity polarities.

3rd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2020.

Ruhaila Maskat and Yuda Munarko. 2019. A taxonomy of Malay social media text. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(1):465–472.

Ruhaila Maskat and Nurazzah Abdul Rahman. 2020. Categorization of Malay social media text and normalization of spelling variations and vowel-less words. *International Journal on Advanced Science, Engineering and Information Technology*, 10(4):1380–1386.

Mohd Masnizah, Fauzi Wan Fariza Paizi, and Amri Jasin. 2018. Teknik pengukuhan perangkak tumpuan melalui modul pengesan bahasa bagi capaian web Bahasa Melayu (Focused crawler enhancement technique with language detection module for Malay web retrieval). *GEMA Online Journal of Language Studies*, 18(3):170–185.

Eleanor Mattern. 2022. The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management. In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Muhammad Rizki Aulia Rahman Maulana and Mohamad Ivan Fanany. 2018a. Indonesian audio-visual speech corpus for multimodal automatic speech recognition. *9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018-January:381–385.

Muhammad Rizki Aulia Rahman Maulana and Mohamad Ivan Fanany. 2018b. Sentence-level Indonesian lip reading with spatiotemporal cnn and gated rnn. *9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018-January:375–380.

Nur Maulidiah Elfajr and Riyanarto Sarno. 2018. Sentiment analysis using weighted emoticons and SentiWordNet for Indonesian language. *3rd International Seminar on Application for Technology of Information and Communication, iSemantic 2018*, pages 234–238.

Candy Olivia Mawalim, Dessi Puji Lestari, and Ayu Purwarianti. 2017. Rule-based reordering and post-processing for Indonesian-Korean statistical machine translation. Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, pages 287–295. The National University (Phillippines).

Zara Maxwell-Smith. 2021. Fossicking in dominant language teaching: Javanese and Indonesian 'low' varieties in language teaching resources. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:24–32.

Zara Maxwell-Smith and Ben Foley. 2021. Developing ASR for Indonesian-English bilingual language teaching. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 131–132.

Zara Maxwell-Smith, Simón González Ochoa, Ben Foley, and Hanna Suominen. 2020. Applications of natural language processing in bilingual language teaching: An Indonesian-English case study. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–134.

Dian Sa'adillah Maylawati, Cecep Nurul Alam, Muhammad Fakhri Muharram, Muhammad Ali Ramdhani, Abdusy Syakur Amin, and Hilmi Aulawi. 2020. The purpose of bellman-ford algorithm to summarize the multiple scientific Indonesian journal articles. *6th International Conference on Wireless and Telematics, ICWT 2020*.

Dian Sa'adillah Maylawati, Yogan Jaya Kumar, Fauziah Binti Kasmin, and Basit Raza. 2019. Sequential pattern mining and deep learning to enhance readability of Indonesian text summarization. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(6):3147–3159.

Niknik Mediyawati, Julio Christian Young, and Samiaji Bintang Nusantara. 2021. U-tapis: Automatic spelling filter as an effort to improve Indonesian language competencies of journalistic students. *Cakrawala Pendidikan*, 40(2):402–412.

Mirwan, Aryo Nugroho, Ferial Hendarta, Rumaisah Hidayatillah, Firdaus Hassan, and Kristovel Printo Nana. 2018. Virtual assistant using lstm networks in Indonesian. *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, pages 652–655.

Paramita Mirza. 2016. Recognizing and normalizing temporal expressions in Indonesian texts. *14th International Conference of the Pacific Association for Computaitonal Linguistics, PACLING 2015*, 593:135–147.

Vivensius Mitra, Herry Sujaini, and Arif Bijaksana Putra Negara. 2017. Rancang bangun aplikasi web scraping untuk korpus paralel indonesia-inggris dengan metode html dom. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, 5(1):36–41.

David Moeljadi and Francis Bond. 2016. Identifying and exploiting definitions in wordnet bahasa. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 227–233. Global Wordnet Association.

Abdul Karim Mohamad, Mailasan Jayakrishnan, and Nurnajwa Hazwani Nawi. 2020a. Classification of Twitter data by sentiment analysis in the Malay language. *International Journal of Emerging Trends in Engineering Research*, 8(6):2730–2738.

Abdul Karim Mohamad, Mailasan Jayakrishnan, and Nurnajwa Hazwani Nawi. 2020b. Employ Twitter data to perform sentiment analysis in the Malay language. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2):1404–1412.

Hasnah Mohamad, Noorul Khairien Abdul Malek, and Noor Husna Abd. Razak. 2020c. Formation of health science terminology by users in general Malay language texts. *GEMA Online Journal of Language Studies*, 20(3):96–112.

Noor Hasnoor Mohamad Nor, Eizah Mat Hussain, and Ahmad Ramizu Abdullah. 2019. Politeness in communication through local children's animated film. *Jurnal Komunikasi: Malaysian Journal of Communication*, 35(4):368–385.

Hassan Mohamed, Nazlia Omar, and Mohd Juzaiddin Ab Aziz. 2018. The effectiveness of using Malay affixes for handling unknown words in unsupervised hmm pos tagger. *International Journal of Engineering & Technology*, 7(4.29):9–12.

Haslizatul Mohamed Hanum and Zainab Abu Bakar. 2016. Detection of Malay phrase breaks using energy and duration. *International Journal of Simulation: Systems, Science and Technology*, 17(32).

Saif M. Mohammad. 2020. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Syuhairah Rahifah Mohammad Najib, Nurazzah Abd Rahman, Normaly Kamal Ismail, Nursyahidah Alias, Zulhilmi Mohamed Nor, and Muhammad Nazir Alias. 2017. Comparative study of machine learning approach on Malay translated hadith text classification based on sanad. *8th International Conference on Mechanical and Manufacturing Engineering, ICME 2017*, 135.

Mohd Amin Mohd Yunus, Aida Mustapha, Rizwan Iqbal, and Noor Azah Samsudin. 2017. An ontological approach towards dialogue based information visualization system: Quran corpus for Juz' Amma. *8th International Conference on Mechanical and Manufacturing Engineering, ICME 2017*, 135.

Nazri Mohd Zakree Ahmad, Kurniawan Tri Basuki, Hamdan Abdul Razak, Salwani Abdullah, and Mohammed Azlan Mis. 2018. Pembangunan taksonomi dari teks Melayu menggunakan algoritma kunang-kunang pembahagi dua sama (Taxonomy development from Malay text using firefly bisection algorithm). *GEMA Online Journal of Language Studies*, 18(2):182–201.

Rosmayati Mohemad, Nazratul Naziah Mohd Muhait, Noor Maizura Mohamad Noor, and Zulaiha Ali Othman. 2020a. A review of named entity recognition and classification on unstructured Malay data. *Journal of Theoretical and Applied Information Technology*, 98(23):3741–3756.

Rosmayati Mohemad, Nazratul Naziah Mohd Muhait, Noor Maizura Mohamad Noor, and Zulaiha Ali Othman. 2020b. Technique on Malay text summarization: A review. *International Journal of Advanced Science and Technology*, 29(6 Special Issue):814–822.

Rosmayati Mohemad, Muhait Nazratul Naziah Mohd, Noor Noor Maizura Mohamad, and Othman Zulaiha Ali. 2020c. Unstructured Malay text analytics model in crime. *IOP Conference Series. Materials Science and Engineering*, 769(1).

Itaza Afiani Mohtar, Masurah Mohamad, Puteri Nursyawati Azzuri, Nurazzah Abd Rahman, and Saidi Adnan Md Nor. 2021. Development of a web-based Jahai-Malay language repository. In *5th International Conference on Information Retrieval and Knowledge Management, CAMP 2021*, pages 14–18. Institute of Electrical and Electronics Engineers Inc.

Putra Fissabil Muhammad, Retno Kusumaningrum, and Adi Wibowo. 2021. Sentiment analysis using word2vec and long short-term memory (lstm) for Indonesian hotel reviews. *5th International Conference on Computer Science and Computational Intelligence, ICCSCI 2020*, 179:728–735.

Muljono, Umriya Afini, and Catur Supriyanto. 2017a. Morphology analysis for hidden markov model based Indonesian part-of-speech tagger. *1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018-January:237–240.

Muljono, Agus Harjoko, Sri Winarsih Nurul Anisa, and Catur Supriyanto. 2020. An evaluation of sentence selection methods on the different phone-sized units for constructing Indonesian speech corpus. *International Journal of Speech Technology*, 23(1):141–147.

Muljono, Surya Sumpeno, Dhany Arifianto, Kiyoaki Aikawa, and Mauridhi Purnomo. 2016. Developing an online self-learning system of Indonesian pronunciation for foreign learners. *International Journal of Emerging Technologies in Learning*, 11(4):83–89.

Muljono, Askarya Qaulan Syadida, De Rosal Ignatius Moses Setiadi, and A. Setyono. 2017b. Sphinx4 for Indonesian continuous speech recognition system. *2017 International Seminar on Application for Technology of Information and Communication, iSemantic 2017*, 2018-January:264–267.

Muljono Muljono, Umriya Afini, Catur Supriyanto, and Raden Nugroho. 2017c. The development of Indonesian POS tagging system for computer-aided independent language learning. *International Journal of Emerging Technologies in Learning*, 12(11):138–150.

Devi Munandar, Endang Suryawati, Dianadewi Riswantini, Achmad Fatchuttamam Abka, Rini Wijayanti, and Andria Arisal. 2017. Pos-tagging for non-English tweets: An automatic approach: (study in Bahasa Indonesia). *1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, 2018-January:219–224.

Yohei Murakami. 2019. Indonesia language sphere: An ecosystem for dictionary development for low-resource languages. *Journal of Physics: Conference Series*, 1192(1).

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman, and Rabiah Abdul Kadir. 2017. English and Malay cross-lingual sentiment lexicon acquisition and analysis. *8th International Conference on Information Science and Applications, ICISA 2017*, 424:467–475.

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman, and Rabiah Abdul Kadir. 2018. A review on building bilingual comparable corpora for resource-limited languages. *4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences, CAMP 2018*, pages 113–118.

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman, and Rabiah Abdul Kadir. 2019. A framework for English and Malay cross-lingual document alignment method. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(1.3 S1):190–195.

Muhammad Zahier Nasrudin, Ruhaila Maskat, and Ramli Musa. 2019. Detecting candidates of depression, anxiety and stress through Malay-written tweets: A preliminary study. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(2):787–793.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2019. Designing a collaborative process to create bilingual dictionaries of Indonesian ethnic languages. *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 3397–3404.

Halim Nataprawira and Michael D. Carey. 2020. Towards developing colloquial Indonesian language pedagogy: A corpus analysis. *Indonesian Journal of Applied Linguistics*, 10(2):382–396.

Rizka Putri Nawangsari, Retno Kusumaningrum, and Adi Wibowo. 2019. Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study. *4th International Conference on Computer Science and Computational Intelligence, ICCSCI 2019*, 157:360–366.

Shekhar Nayak, C Shiva Kumar, G. Ramesh, Saurabhchand Bhati, and K. Sri Rama Murty. 2019. Virtual phone discovery for speech synthesis without text. *7th IEEE Global Conference on Signal and Information Processing, GlobalSIP 2019*.

Bagas Pradipabista Nayoga, Ryan Adipradana, Ryan Suryadi, and Derwin Suhartono. 2021. Hoax analyzer for Indonesian news using deep learning models. *5th International Conference on Computer Science and Computational Intelligence, ICCSCI 2020*, 179:704–712.

Wayan Suastini Ni, Ketut Artawa, Yadnya Ida Bagus Putra, and I. Ketut Darma Laksana. 2018. Translation and markedness. *International Journal of Comparative Literature & Translation Studies*, 6(4):28–32.

Annisa Maulida Ningtyas and Guntur Budi Herwanto. 2018. The influence of negation handling on sentiment analysis in Bahasa Indonesia. *5th International Conference on Data and Software Engineering, ICoDSE 2018*.

Made Nindyatama Nityasya, Rahmad Mahendra, and Mirna Adriani. 2019. Hypernym-hyponym relation extraction from Indonesian wikipedia text. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 285–289.

Muhammad Nizami and Ayu Purwarianti. 2017. Modification of chu-liu/edmonds algorithm and mira learning algorithm for dependency parser on Indonesian language. *2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*.

Shahrul Azman Mohd Noah, Nazlena Mohamad Ali, and Mohd Sabri Hasan. 2018. Generation of news headline for Malay language based on term features. *GEMA Online Journal of Language Studies*, 18(4):42–59.

Zakiah Noh, Siti Zaleha Zainal Abidin, and Nasiroh Omar. 2019. Poetry visualization in digital technology. *Knowledge Management and Organizational Learning*, 7:171–195.

Hiroki Nomoto. 2020. Towards genuine stemming and lemmatization in Malay/Indonesian. In *Proceedings of the 26th Annual Meeting of the Natural Language Processing Society (March 2020)*.

Hiroki Nomoto, Shiro Akasegawa, and Asako Shiohara. 2018a. Reclassification of the leipzig corpora collection for Malay and Indonesian. *Nusa*, 65:47–66.

Hiroki Nomoto and David Moeljadi. 2019. Linguistic studies using large annotated corpora: introduction. *Nusa*, pages 1–6.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018b. Tufs Asian language parallel corpus (talpco). *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, pages 436–439.

Huzaimi Karimah Mohd Noor Noor, Shahrul Azman Mohd Noah, and Mohd Juzaiddin Ab Aziz. 2020. Classification of short possessive clitic pronoun nya in Malay text to support anaphor candidate determination. *Journal of Information and Communication Technology*, 19(4):513–532.

Noorhuzaimi Moh Noor, Junaida Sulaiman, and Shahrul Azman Noah. 2016. Malay name entity recognition using limited resources. *Advanced Science Letters*, 22(10):2968–2971.

Muhamad Nor Azlizawati Binti, Norisma Idris, and Saloot Mohammad Arshi. 2017. Proposal: A hybrid dictionary modelling approach for Malay tweet normalization. *Journal of Physics: Conference Series*, 806(1).

Nor Nor Fariza Mohd, Rahman Anis Nadiah Che Abdul, Azhar Jaluddin, Abdullah Imran Ho, and Sabrina Tiun. 2019. A corpus driven analysis of representations around the word 'ekonomi' in Malaysian hansard corpus. *GEMA Online Journal of Language Studies*, 19(4):66–95.

Awal Norsimah Mat, Azhar Jaludin, Rahman Anis Nadiah Che Abdul, and Imran Ho-Abdullah. 2019. "Is Selangor in Deep Water?": A corpus-driven account of air/water in the Malaysian hansard corpus (mhc). *GEMA Online Journal of Language Studies*, 19(2):99–120.

Sashi Novitasari, Dessi Puji Lestari, Sakriani Sakti, and Ayu Purwarianti. 2019. Rude-words detection for Indonesian speech using support vector machine. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 19–24.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138.

Aditya Alif Nugraha, Anditya Arifianto, and Suyanto. 2019. Generating image description on Indonesian language using convolutional neural network and gated recurrent unit. *7th International Conference on Information and Communication Technology, ICoICT 2019*.

Yanuar Nurdiansyah, Saiful Bukhori, and Rahmad Hidayat. 2018. Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method. *Journal of Physics: Conference Series*, 1008(1).

M. Nurilman Baehaqi, Med. Irzal, and Fariani Hermin Indiyah. 2019. Morphological analysis of speech translation into Indonesian sign language system (SIBI) on android platform. *11th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019*, pages 205–210.

Vessa Rizky Oktavia, Umi Laili Laili Yuhana, Chastine Fatichah, and Ayu Purwarianti. 2021. WPS: Application for generating answer of word problem in Bahasa Indonesia. In *8th International Conference on ICT for Smart Society, ICISS 2021*. Institute of Electrical and Electronics Engineers Inc.

Ikmi Nur Oktavianti. 2019. A corpus-based analysis of English core modal verbs and their counterparts in Indonesian. *International Journal of Scientific and Technology Research*, 8(12):2811–2819.

Ikmi Nur Oktavianti and Zanuar Anggun Pramesti. 2019. Frequency of verbs in lifestyle column in the Jakarta Post and the relation to text characteristics: A corpus-based analysis. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature*, 7(2):233–246.

Nasiroh Omar, Ahmad Farhan Hamsani, Nur Atiqah Sia Abdullah, and Siti Zaleha Zainal Abidin. 2017. Construction of Malay abbreviation corpus based on social media data. *Journal of Engineering and Applied Sciences*, 12(3):468–474.

Salehah Omar, Juhaida Abu Bakar, Maslinda Mohd Nadzir, Nor Hazlyna Harun, and Nooraini Yusoff. 2021. Text simplification for Malay corpus: A review. In *6th International Conference on Computer and Information Sciences, ICCOINS 2021*, pages 345–350. Institute of Electrical and Electronics Engineers Inc.

Veronica Ong, Anneke Dwi Sesarika Rahmanto, Williem, Derwin Suhartono, Aryo E. Nugroho, Esther W. Andangsari, and Muhamad N. Suprayogi. 2017. Personality prediction based on Twitter information in Bahasa Indonesia. *2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, pages 367–372.

Veronica Ong, Anneke Dwi Sesarika Rahmanto, Williem Williem, Nicholaus H. Jeremy, Derwin Suhartono, and Esther W. Andangsari. 2021. Personality modelling of Indonesian Twitter users with xgboost based on the five factor model. *International Journal of Intelligent Engineering and Systems*, 14(2):248–261.

Norzanita Othman, Nor Nor Fariza Mohd, and Noraini Ibrahim. 2019. Linguistic representation of violence in judicial opinions in Malaysia. *GEMA Online Journal of Language Studies*, 19(2):82–98.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Anea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, 10(1):89–89.

Endang Wahyu Pamungkas and Divi Galih Prasetyo Putri. 2017. An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia. *6th International Annual Engineering Seminar, InAES 2016*, pages 28–31.

Ningrum Panggih Kusuma, Tatdow Pansombut, and Attachai Ueranantasun. 2020. Text mining of online job advertisements to identify direct discrimination during job hunting process: A case study in Indonesia. *PLoS ONE*, 15(6):e0233746.

Bagus Pragnya Paramarta. 2018. Analisis korpus terhadap idiom Bahasa Indonesia yang berbasis nama binatang. *Lingua*, 14(1):18–25.

Edwina Anky Parande and Suyanto Suyanto. 2019. Indonesian graphemic syllabification using a nearest neighbour classifier and recovery procedure. *International Journal of Speech Technology*, 22(1):13–20.

Jasman Pardede and Mira Musrini Barmawi. 2016. Implementation of lsi method on information retrieval for text document in Bahasa Indonesia. *Internetworking Indonesia Journal*, 8(1):83–87.

W. G. S. Parwita. 2020. A document recommendation system of stemming and stopword removal impact: A web-based application. *Journal of Physics: Conference Series*, 1469(1).

Fahmi Candra Permana, Yusep Rosmansyah, and Atje Setiawan Abdullah. 2017. Naive bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter social media. *Asian Mathematical Conference 2016, AMC 2016*, 893.

Micah D.J. Peters, Christina M. Godfrey, Hanan Khalil, Patricia McInerney, Deborah Parker, and Cassia Baldini Soares. 2015. Guidance for conducting systematic scoping reviews. *JBI Evidence Implementation*, 13(3):141–146.

Mai T. Pham, Andrijana Rajić, Judy D. Greig, Jan M. Sargeant, Andrew Papadopoulos, and Scott A. McEwen. 2014. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research synthesis methods*, 5(4):371–385.

Yuen Chi Phang, Azleena Mohd Kassim, and Ernest Mangantig. 2021. Concerns of thalassemia patients, carriers, and their caregivers in Malaysia: Text mining information shared on social media. *Healthcare Informatics Research*, 27(3):200–213.

Yeong-Tsann Phua, Kwang-Hooi Yew, Oi-Mean Foong, and Matthew Yok-Wooi Teow. 2020. Assessing suitable word embedding model for Malay language through intrinsic evaluation. *2020 International Conference on Computational Intelligence, ICCI 2020*, pages 202–210.

Dion Ajie Poetra, Tricya Esterina Widagdo, and Fazat Nur Azizah. 2019. Natural language interface to database (NLIDB) for query with temporal aspect. *2019 International Conference on Data and Software Engineering, ICoDSE 2019*.

Faizal Adhitama Prabowo, Muhammad Okky Ibrohim, and Indra Budi. 2019. Hierarchical multi-label classification to identify hate speech and abusive language on Indonesian Twitter. *6th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2019*.

Bayu Trisna Pratama, Ema Utami, and Andi Sunyoto. 2019. A comparison of the use of several different resources on lexicon based Indonesian sentiment analysis on app review dataset. *1st International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pages 282–287.

Septya Egho Pratama, Wahyudin Darmalaksana, Dian Sa'adillah Maylawati, Hamdan Sugilar, Teddy Mantoro, and Muhammad Ali Ramdhani. 2020. Weighted inverse document frequency and vector space model for hadith search engine. *Indonesian Journal of Electrical Engineering and Computer Science*, 18(2):1004–1014.

Timothy Pratama and Ayu Purwarianti. 2017. Topic classification and clustering on Indonesian complaint tweets for Bandung government using supervised and unsupervised learning. *2017 International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2017*.

Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. 2018. Study of hoax news detection using naïve bayes classifier in Indonesian language. *11th International Conference on Information and Communication Technology and System, ICTS 2017*, 2018-January:73–78.

Nur Indah Pratiwi, Indra Budi, and Ika Alfina. 2019. Hate speech detection on Indonesian Instagram comments using fasttext approach. *10th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, pages 447–450.

Prihantoro. 2016. The influence of students' L1 and spoken English in English writing: A corpus-based research. *TEFLIN Journal*, 27(2):217–245.

Prihantoro. 2021. An evaluation of morphind's morphological annotation scheme for Indonesian. *Corpora*, 16(2):287.

D. Purnamasari, R. Arianty, D. T. Susetianingtias, and R. D. Kusumawati. 2016. Query rewriting and corpus of semantic similarity as encryption method for documents in Indonesian language. *2nd International Conference on Electrical Systems, Technology and Information, ICESTI 2015*, 365:565–571.

K. K. Purnamasari and I. S. Suwardi. 2018. Rule-based part of speech tagger for Indonesian language. *International Conference on Informatics, Engineering, Science and Technology, INCITEST 2018*, 407.

Yohanes Sigit Purnomo W.P, Yogan Jaya Kumar, and Nur Zareen Zulkarnain. 2020. Understanding quotation extraction and attribution: towards automatic extraction of public figure's statements for journalism in Indonesia. *Global Knowledge, Memory and Communication*.

Dewi Puspita and Kamal Yusuf. 2020. Sketching the semantic change of jahanam and hijrah: A corpus based approach to manuscripts of Arabic-Indonesian lexicon. *Arabi: Journal of Arabic Studies*, 5(1):1–10.

Nurnasran Puteh, Mohd Zabidin Husin, Hatim Mohamad Tahir, and Azham Hussain. 2019. Building a question classification model for a Malay question answering system. *International Journal of Innovative Technology and Exploring Engineering*, 8(5s):184–190.

I Gede Manggala Putra and Dade Nurjanah. 2020. Hate speech detection in Indonesian language Instagram. *12th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, pages 413–420.

M. Iqbal D. Putra, Budi Irmawati, Wirarama Wedashwara, Dita Pramesti, and Siti Oryza Khairunnisa. 2020. Age group based document classification in Bahasa Indonesia. *2020 International Conference on Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2020*.

Rahardyan Bisma Setya Putra, Ema Utami, and Suwanto Raharjo. 2018a. Non-formal affixed word stemming in Indonesian language. *1st International Conference on Information and Communications Technology, ICOIACT 2018*, 2018-January:531–536.

Rahardyan Bisma Setya Putra, Ema Utami, and Suwanto Raharjo. 2019. Accuracy measurement on Indonesian non-formal affixed word stemming with levenhstein. *2nd International Conference on Information and Communications Technology, ICOIACT 2019*, pages 486–490.

Syopiansyah Jaya Putra, Muhamad Nur Gunawan, Ismail Khalil, and Teddy Mantoro. 2017. Sentence boundary disambiguation for Indonesian language. *19th International Conference on Information Integration and Web-Based Applications and Services, iiWAS2017*, pages 587–590.

Syopiansyah Jaya Putra, Ismail Khalil, Muhamad Nur Gunawan, Riva'l Amin, and Tata Sutabri. 2018b. A hybrid model for social media sentiment analysis for Indonesian text. *20th International Conference on Information Integration and Web-Based Applications and Services, iiWAS 2018*, pages 297–301.

Fanda Yuliana Putri, Devin Hoesen, and Dessi Puji Lestari. 2019a. Rule-based pronunciation models to handle oov words for Indonesian automatic speech recognition system. *5th International Conference on Science in Information Technology, ICSITech 2019*, pages 246–251.

S. K. Putri, A. Amalia, E. B. Nababan, and O. S. Sitompul. 2021. Bahasa Indonesia pre-trained word vector generation using word2vec for computer and information technology field. In *5th International Conference on Computing and Applied Informatics, ICCAI 2020*, volume 1898. IOP Publishing Ltd.

Wahyuningdiah Trisari Harsanti Putri, Muhammad Singgih Prastio, Retno Hendrowati, Yustiana Sari, and Harry Tursulistyono Yani Achsan. 2019b. Content-based filtering model for recommendation of Indonesian legal article study case of klinik hukumonline. *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, pages 9–14.

Xin Ying Qiu and Gangqin Zhu. 2016. Learning Indonesian-Chinese lexicon with bilingual word embedding models and monolingual signals. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 188–193.

Valdi Rachman, Rahmad Mahendra, Alfan Farizki Wicaksono, Ahmad Rizqi Meydiarso, and Fariz Ikhwantri. 2018a. Semantic role labeling in conversational chat using deep bi-directional long short-term memory networks with attention mechanism. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics.

Valdi Rachman, Septiviana Savitri, Fithriannisa Augustianti, and Rahmad Mahendra. 2018b. Named entity recognition on Indonesian Twitter posts using long short-term memory networks. *9th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018-January:228–232.

Suwanto Raharjo, Retantyo Wardoyo, and Agfianto E. Putra. 2018. Rule based sentence segmentation of Indonesian language. *Journal of Engineering and Applied Sciences*, 13(21):8986–8992.

Suwanto Raharjo, Retantyo Wardoyo, and Agfianto E. Putra. 2020. Detecting proper nouns in Indonesian-language translation of the Quran using a guided method. *Journal of King Saud University - Computer and Information Sciences*, 32(5):583–591.

Anis Nadiah Che Abdul Rahman, Imran Ho Abdullah, Intan Safinaz Zainudin, Sabrina Tiun, and Azhar Jaludin. 2021a. Domain-specific stop words in Malaysian parliamentary debates 1959 - 2018. *GEMA Online Journal of Language Studies*, 21(2):1–27.

Arief Rahman. 2018. Medical named entity recognition for Indonesian language using word representations. *IOP Conference Series. Materials Science and Engineering*, 325(1).

Arief Rahman and Ayu Purwarianti. 2017. Ensemble technique utilization for Indonesian dependency parser. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 64–71.

Nurazzah Abd Rahman, Faiz Ikhwan Mohd Rafhan Syamil, and Shaiful Bakhtiar Bin Rodzman. 2020. Development of mobile application for Malay translated hadith search engine. *Indonesian Journal of Electrical Engineering and Computer Science*, 20(2):932–938.

Nurazzah Abd Rahman, Siti Nur Afiqah Ramlam, Natasha Aleza Azhar, Haslizatul Mohamed Hanum,

Noor Ida Ramli, and Najahudin Lateh. 2021b. Automatic text summarization for Malay news documents using latent dirichlet allocation and sentence selection algorithm. In *5th International Conference on Information Retrieval and Knowledge Management, CAMP 2021*, pages 36–40. Institute of Electrical and Electronics Engineers Inc.

Nurazzah Abd Rahman, Afiqah Bazlla Md Soom, and Normaly Kamal Ismail. 2017. Enhancing latent semantic analysis by embedding tagging algorithm in retrieving Malay text documents. *Studies in Computational Intelligence*, 710:309–319.

Rinaldi Andrian Rahmanda, Mirna Adriani, and Dipta Tanaya. 2019. Cross language information retrieval using parallel corpus with bilingual mapping method. *23rd International Conference on Asian Language Processing, IALP 2019*, pages 222–227.

Muhammad Abdillah Rahmat, Indrabayu, and Intan Sari Areni. 2019. Hoax web detection for news in bahasa using support vector machine. *2nd International Conference on Information and Communications Technology, ICOIACT 2019*, pages 332–336.

Laili Etika Rahmawati, Anggi Niasih, Hari Kusmanto, and Harun Joko Prayitno. 2020. Environmental awareness content for character education in grade 10 in Indonesian language student textbooks. *International Journal of Innovation, Creativity and Change*, (4):161–174.

F. Rahutomo, A. A. Septarina, M. Sarosa, A. Setiawan, and M. M. Huda. 2019. A review on Indonesian machine translation. *4th Annual Applied Science and Engineering Conference, AASEC 2019*, 1402.

Faisal Rahutomo, Alfi Samudro Mulyo, and Pramana Yoga Saputra. 2018. Automatic grammar checking system for Indonesian. *1st International Conference on Applied Science and Technology, iCAST 2018*, pages 308–313.

Reza Rahutomo, Arif Budiarto, Kartika Purwandari, and Anzaludin Samsinga Perbangsa. 2020. Ten-year compilation of #savekpk Twitter dataset. *5th International Conference on Information Management and Technology, ICIMTech 2020*, pages 185–190.

Roza Athirah Raja, Soon Lay-Ki, and Haw Su-Cheng. 2019. Exploring edit distance for normalising out-of-vocabulary Malay words on social media. *MATEC Web of Conferences*, 255:03001.

Gede Primahadi Wijaya Rajeg. 2020. Linguistik korpus kuantitatif dan kajian semantik leksikal sinonim emosi Bahasa Indonesia. *Linguistik Indonesia*, 38(2):123–150.

Gede Primahadi Wijaya Rajeg, Karlina Denistia, and Simon Musgrave. 2019. Vector space models and the usage patterns of Indonesian denominal verbs: a case study of verbs with meN-, meN-/-kan, and meN-/-i affixes. *Nusa*, pages 35–76.

Rajesvary Rajoo and Ching Chee Aun. 2016. Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages. *2016 IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2016*, pages 35–39.

Nahda Rosa Ramadhanti and Siti Mariyah. 2019. Document similarity detection using Indonesian language word2vec model. *3rd International Conference on Informatics and Computational Sciences, ICICOS 2019*.

Muhammad Ali Ramdhani, Dian Sa'adillah Maylawati, and Teddy Mantoro. 2020. Indonesian news classification using convolutional neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(2):1000–1009.

Al-khulaidi Rami Ali and Akmeliawati Rini. 2017. Speech to text translation for Malay language. *IOP Conference Series. Materials Science and Engineering*, 260(1).

I. Ramli, N. Jamil, N. Seman, and N. Ardi. 2017. The first Malay language storytelling text-to-speech (TTS) corpus for humanoid robot storytellers. *Journal of Fundamental and Applied Sciences*, 9(4S):340–354.

Izzad Ramli, Jamil Nursuriati, and Noraini Seman. 2021. An iterated two-step sinusoidal pitch contour formulation for expressive speech synthesis. *Journal of Information and Communication Technology*, 20(4):489–510.

Izzad Ramli, Noraini Seman, Norizah Ardi, and Nursuriati Jamil. 2016. Prosody analysis of Malay language storytelling corpus. *18th International Conference on Speech and Computer, SPECOM 2016*, 9811 LNCS:563–570.

Bali Ranaivo-Malançon, Suhaila Saee, Rosita Mohamed Othman, and Jennifer Fiona Wilfred Busu. 2017. Transforming semi-structured indigenous dictionary into machine-readable dictionary. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(3-11):7–11.

Francisco Rangel, Paolo Rosso, Julian Brooke, and Alexandra L Uitdenbogerd. 2018. Cross-corpus native language identification via statistical embedding. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 39–43.

Ihda Rasyada, Yuliana Setiowati, Aliridho Barakbah, and M. Tafaquh Fiddin Al Islami. 2020. Sentiment analysis of bpjs kesehatan's services based on affective models. *2020 International Electronics Symposium, IES 2020*, pages 549–556.

Anak Agung Putri Ratna, Randy Sanjaya, Tomi Wirianata, and Prima Dewi Purnamasari. 2017. Word level auto-correction for latent semantic analysis

based essay grading system. *15th International Conference on Quality in Research: International Symposium on Electrical and Computer Engineering, QiR 2017*, 2017-December:235–240.

Anak Agung Putri Ratna, Naiza Astri Wulandari, Aaliyah Kaltsum, Ihsan Ibrahim, and Prima Dewi Purnamasari. 2019. Answer categorization method using k-means for Indonesian language automatic short answer grading system based on latent semantic analysis. *16th International Conference on Quality in Research, QIR 2019*.

Sitti Munirah Abdul Razak, Muhamad Sadry Abu Seman, Wan Ali, Wan Yusoff Wan, Noor Hasrul Nizan, and Mohammad Noor. 2019. Malay manuscripts transliteration using statistical machine translation (SMT). *1st International Conference on Artificial Intelligence and Data Sciences, AiDAS 2019*, pages 137–141.

Sitti Munirah Abdul Razak, Muhamad Sadry Abu Seman, Wan Ali Wan Yusoff Wan Mamat, and Noor Hasrul Nizan Mohammad Noor. 2018. Transliteration engine for union catalogue of Malay manuscripts in Malaysia: E-JAWI version 3. *2018 International Conference on Information and Communication Technology for the Muslim World, ICT4M 2018*, pages 58–63.

Husna Faredza Mohamed Redzwan, Khairul Azam Bahari, Anida Sarudin, and Zulkifli Osman. 2020. Strategi pengukuran upaya berbahasa menerusi kesantunan berbahasa sebagai indikator profesionalisme guru pelatih berasaskan skala morfofonetik, sosiolinguistik dan sosiopragmatik (Linguistic politeness as an indicator of trainee teacher professionalism: A language abilty measurement strategy based on morphophonetic, sociolinguistic and sociopragmatic scales). *Malaysian Journal of Learning and Instruction*, 17(1):213–254.

Rianto, Achmad Benny Mutiara, Eri Prasetyo Wibowo, and Paulus Insap Santosa. 2021. Improving stemming techniques for non-formal Indonesian sentences using incorbiz. *ICIC Express Letters*, 15(1):67–74.

Mohd Arizal Shamsil Mat Rifin and Mohd Pouzi Hamzah. 2017. Incorporating knowledge base in unsupervised approach of word sense disambiguation of Malay documents. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(3-4 Special Issue):119–122.

Muhammad Rif'at, Rahmad Mahendra, Indra Budi, and Haryo Akbarianto Wibowo. 2018. Towards product attributes extraction in Indonesian e-commerce platform. *Computacion y Sistemas*, 22(4):1367–1375.

Dewi Riyanti, M. Arif Bijaksana, and Adiwijaya. 2018. Automatic semantic orientation of adjectives for Indonesian language using pmi-ir and clustering. *International Conference on Data and Information Science 2017, ICoDIS 2017*, 971.

Faizal Riza, Saefulloh Rifai, Akmal Dirgantara, Sfenrianto, Rasenda, and Syarifudin Herdyansyah. 2020. Information retrieval technique for Indonesian pdf document with modified stemming porter method using php. *Journal of Physics: Conference Series*, 1477(3).

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái, Vichet Chea, Vichet Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2017. Introduction of the asian language treebank. *19th Annual Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA 2016*, pages 1–6.

Arra'Di Nur Rizal and Sara Stymne. 2020. Evaluating word embeddings for Indonesian–English code-mixed text based on synthetic data. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 26–35. European Language Resources Association.

Afian Syafaadi Rizki, Aris Tjahyanto, and Rahmat Trialih. 2019. Comparison of stemming algorithms on Indonesian text processing. *TELKOMNIKA*, 17(1):95–102.

Shaiful Bakhtia Rodzman, Sumayyah Hasbullah, Nomaly Kamal Ismail, Nurazzah Abd Rahman, Zulhilmi Mohamed Nor, and Ahmad Yunus Mohd Noor. 2020. Fabricated and Shia Malay translated hadith as negative fuzzy logic ranking indicator on Malay information retrieval. *ASM Science Journal*, 13(Special Issue 3):101–108.

Shaiful Bakhtia Rodzman, Nomaly Kamal Ismail, Nurazzah Abd Rahman, Syed Ahmad Aljunid, Zulhilmi Mohamed Nor, and Ahmad Yunus Mohd Noor. 2019. Domain specific concept ontologies and text summarization as hierarchical fuzzy logic ranking indicator on Malay text corpus. *Indonesian Journal of Electrical Engineering and Computer Science*, 15(3):1527–1534.

Shaiful Bakhtia Rodzman, Mohamad Fitri Izuan Abdul Ronie, Normaly Kamal Ismail, Nurazzah Abd Rahman, Fatimah Ahmad, and Zulhilmi Mohamed Nor. 2018. Analyzing Malay stemmer performance towards fuzzy logic ranking function on Malay text corpus. *4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences, CAMP 2018*, pages 36–41.

Ade Romadhony, Ayu Purwarianti, and Dwi Hendratmo Widyantoro. 2018. Rule-based Indonesian open information extraction. *5th International Conference on Advanced Informatics: Concepts Theory and Applications, ICAICTA 2018*, pages 107–112.

Fadhilah Rosdi, Mumtaz Begum Mustafa, and Siti Salwah Salim. 2017. Assessing automatic speech recognition in measuring speech intelligibility: A study of

Malay speakers with speech impairments. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–6.

Rosiyana Rosiyana. 2020. Pengajaran bahasa dan pemerolehan bahasa kedua dalam pembelajaran BIPA (Bahasa Indonesia Penutur Asing). *Jurnal Ilmiah KORPUS*, 4(3):374–382.

Suhanah Rosnan, Nurazzah Abd Rahman, Shahirah Mohamed Hatim, and Zahirah Hamid Ghul. 2019. Performance evaluation of inverted files, b-tree and b+ tree indexing algorithm on Malay text. *4th International Conference and Workshops on Recent Advances and Innovations in Engineering, ICRAIE 2019*.

Raphael Rubino, Benjamin Marie, Raj Dabre, Atushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2020. Extremely low-resource neural machine translation for asian languages. *Machine Translation*, 34(4):347–382.

Fitrah Rumaisa, Halizah Basiron, Zurina Saaya, and Noorli Khamis. 2019. Development of multilingual social media data corpus: Development and evaluation. *International Journal of Innovation, Creativity and Change*, 6(5):1–14.

Fitrah Rumaisa, Halizah Basiron, Zurina Saaya, and Yoki Muchsam. 2020. BMBI: A development of a special corpus on homonyms for multi-lingual sentiment analysis. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4).

Andre Rusli, Julio Christian Young, and Ni Made Satvika Iswari. 2020. Identifying fake news in Indonesian via supervised binary text classification. *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2020*, pages 86–90.

Tatyana Ruzsics and Tanja Samardzic. 2017. Neural sequence-to-sequence learning of internal word structure. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 184–194. Association for Computational Linguistics.

Mastura Md Saad, Nursuriati Jamil, and Raseeda Hamzah. 2018a. Evaluation of support vector machine and decision tree for emotion recognition of Malay folklores. *Bulletin of Electrical Engineering and Informatics*, 7(3):479–486.

Nurui Huda Mohd Saad and Nor Hashimah Jalaluddin. 2020. Imbuhan meN- dengan kata nama konkrit unsur alam: Analisis teori relevans (Prefix meN- with concrete nouns of natural elements: A relevance theory analysis). *GEMA Online Journal of Language Studies*, 20(3):136–155.

Saidah Saad and Mansor Mohamed Kamil. 2018. Pendekatan teknik pengecaman entiti nama bagi capaian berita jenayah Bahasa Melayu (Named entity recognition approach for Malay crime news retrieval). *GEMA Online Journal of Language Studies*, 18(4):216–235.

Suziana Mat Saad, Nor Hashimah Jalaluddin, and Imran Ho-Abdullah. 2018b. Conceptual metaphor and linguistic manifestations in Malay and French: A cognitive analysis. *GEMA Online Journal of Language Studies*, 18(3):114–134.

Johnny Saldaña. 2016. *The Coding Manual for Qualitative Researchers*, third edition edition. SAGE, London; Los Angeles, CA.

Calvin Erico Rudy Salim and Derwin Suhartono. 2021. Long short-term memory for hate speech and abusive language detection on Indonesian YouTube comment section. In *2021 11th International Workshop on Computer Science and Engineering, WCSE 2021*, pages 193–200. International Workshop on Computer Science and Engineering (WCSE).

M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad. 2018. Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(2-7):121–126.

Muhammad Sharilazlan Salleh, Siti Azirah Asmai, Halizah Basiron, and Sabrina Ahmad. 2017. A Malay named entity recognition using conditional random fields. *5th International Conference on Information and Communication Technology, ICoIC7 2017*.

Mohammad Arshi Saloot, Norisma Idris, AiTi Aw, and Dirk Thorleuchter. 2016. Twitter corpus creation: The case of a Malay chat-style-text corpus (MCC). *Digital Scholarship in the Humanities*, 31(2):227–243.

Nur-Hana Samsudin and Lukman Nurhaqi Rahim. 2019. Rapid heteronym disambiguation for text-to-speech system. *4th International Conference and Workshops on Recent Advances and Innovations in Engineering, ICRAIE 2019*.

Lidia Sandra and Ford Lumbangaol. 2021. When homecoming is not coming: 2021 homecoming ban sentiment analysis on Twitter data using support vector machine algorithm. In *8th International Conference on ICT for Smart Society, ICISS 2021*. Institute of Electrical and Electronics Engineers Inc.

Agung Santosa, Andi Djalal Latief, Hammam Riza, Asril Jarin, Lyla Ruslana Aini, Gunarso, Gita Citra Puspita, Muhammad Teduh Uliniansyah, Elvira Nurfadhilah, Harnum A. Prafitia, and Made Gunawan. 2019. The architecture of speech-to-speech translator for mobile conversation. *22nd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2019*.

Joan Santoso, Gunawan Gunawan, Hermes Vincentius Gani, Eko Mulyanto Yuniarno, Mochamad Hariadi, and Mauridhi Hery Purnomo. 2016. Noun phrases extraction using shallow parsing with c4.5 decision tree algorithm for Indonesian language ontology building. *15th International Symposium on Communications and Information Technologies, ISCIT 2015*, pages 149–152.

Joan Santoso, Esther Irawati Setiawan, Christian Nathaniel Purwanto, Eko Mulyanto Yuniarno, Mochamad Hariadi, and Mauridhi Hery Purnomo. 2021. Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory. *Expert Systems with Applications*, 176.

Erikson Saragih, Syahron Lubis, Amrin Saragih, Roswita Silalahi, and M. Hum. 2017. Ideational grammatical metaphors in doctrinal verses of The Bible in Indonesian version. *Theory and Practice in Language Studies*, 7(10):847–854.

Anida Sarudin, Mazura Mastura Muhammad, Muhamad Fadzllah Zaini, Zulkifli Osman, and Muhammad Anas Al Muhsin. 2020a. Collocation analysis of variants of intensifies in classical Malay texts. In *2020 Conference of the Global Council on Anthropological Linguistics in Asia, GLOCAL 2020*, volume 2020-January, pages 352–357. Global Council on Anthropological Linguistics.

Anida Sarudin, Mazura Mastura Muhammad, Muhamad Fadzllah Zaini, Husna Faredza Mohamed Redzwan, and Siti Saniah Abu Bakar. 2020b. The relationship between astronomy and architecture as an element of Malay intelligentsia. In *2020 Conference of the Global Council on Anthropological Linguistics in Asia, GLOCAL 2020*, volume 2020-January, pages 358–363. Global Council on Anthropological Linguistics.

Dhanang Hadhi Sasmita, Alfan Farizki Wicaksono, Samuel Louvan, and Mirna Adriani. 2018. Unsupervised aspect-based sentiment analysis on Indonesian restaurant reviews. *21st International Conference on Asian Language Processing, IALP 2017*, 2018-January:383–386.

Siti Syakirah Sazali, Zainab Abu Bakar, and Jafreezal Jaafar. 2016. Word prediction algorithm in resolving ambiguity in Malay text. *3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, pages 1347–1352.

Siti Syakirah Sazali, Nurazzah Abdul Rahman, and Zainab Abu Bakar. 2017. Information extraction: Evaluating named entity recognition from classical Malay documents. *3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016*, pages 48–53.

Siti Syakirah Sazali, Nurazzah Abdul Rahman, and Zainab Abu Bakar. 2020. Characteristics of Malay translated hadith corpus. *Journal of King Saud University - Computer and Information Sciences*.

Noraini Seman and Ahmad Firdaus Norazam. 2019. Hybrid methods of Brandt's generalised likelihood ratio and short-term energy for Malay word speech segmentation. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(1):283–291.

Amalia Agung Septarina, Faisal Rahutomo, and Moechammad Sarosa. 2019. Machine translation of Indonesian: A review. *Communications in Science and Technology*, 4(1):12–19.

Garin Septian, Ajib Susanto, and Guruh Fajar Shidik. 2017. Indonesian news classification based on nabana. *2017 International Seminar on Application for Technology of Information and Communication, iSemantic 2017*, 2018-January:175–180.

Ali Akbar Septiandri, Yosef Ardhito Winatmoko, and Ilham Firdausi Putra. 2020. Knowing right from wrong: Should we use more complex models for automatic short-answer scoring in Bahasa Indonesia? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 1–7.

Reza Setiabudi, Ni Made Satvika Iswari, and Andre Rusli. 2021. Enhancing text classification performance by preprocessing misspelled words in Indonesian language. *TELKOMNIKA*, 19(4):1234–1241.

Evelyn Setiani and Win Ce. 2018. Text classification services using naïve bayes for Bahasa Indonesia. *3rd International Conference on Information Management and Technology, ICIMTech 2018*, pages 361–366.

Esther Irawati Setiawan, Andy Januar Wicaksono, Joan Santoso, Yosi Kristian, Surya Sumpeno, and Mauridhi Hery Purnomo. 2018. N-gram keyword retrieval on association rule mining for predicting teenager deviant behavior from school regulation. *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018*, pages 325–328.

Roziyani Setik, Raja Mohd Tariqi Raja Lope Ahmad, and Suziyanti Marjudi. 2021. Aspect-based sentiment analysis for posts on friday prayer during mco in Malaysia. In *2021 International Congress of Advanced Technology and Engineering, ICOTEN 2021*. Institute of Electrical and Electronics Engineers Inc.

Roziyani bt Setik, Raja Mohd Tariqi Bin Raja Ahmad, Suziyanti bt Marjudi, Azhar bin Hamid, Wan Hassan Basri bin Wan Ismail, Zuraidy bin Adnan, and Wan Azlan bin Wan Hassan. 2018. Exploiting Malay corpus on islamic issue using sketch engine. *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, pages 281–286.

Yuliana Setiowati, Arif Djunaidy, and Daniel Oranova Siahaan. 2019. Pair extraction of aspect and implicit opinion word based on its co-occurrence in corpus of Bahasa Indonesia. *2nd International Seminar on*

Research of Information Technology and Intelligent Systems, ISRITI 2019, pages 73–78.

Verena Severina and Masayu Leylia Khodra. 2019. Multidocument abstractive summarization using abstract meaning representation for Indonesian language. *2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*.

Noah Shahrul Azman Mohd, Ali Nazlena Mohamad, and Hasan Mohd Sabri. 2018a. Penentuan fitur bagi pengekstrakan tajuk berita akhbar Bahasa Melayu (Determining features of news headline in Malay news document). *GEMA Online Journal of Language Studies*, 18(2):154–167.

Noah Shahrul Azman Mohd, Ali Nazlena Mohamad, and Hasan Mohd Sabri. 2018b. Penjanaan ringkasan isi utama berita Bahasa Melayu berdasarkan ciri kata (Generation of news headline for Malay language based on term features). *GEMA Online Journal of Language Studies*, 18(4):42–60.

Gayane Shalunts, Gerhard Backfried, and Helmy Syakh Alam. 2018. Sentiment analysis in Indonesian and French by sentisail. *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, 2018-February:69–75.

Nurul Fathiyah Shamsudin, Halizah Basiron, and Zurina Sa'aya. 2016. Lexical based sentiment analysis - verb, adverb & negation. *Journal of Telecommunication, Electronic and Computer Engineering*, 8(2):161–166.

Asako Shiohara, Yuta Sakon, and Hiroki Nomoto. 2019. Discourse functions of the two non-active voices in Indonesian: based on the web corpus in MALINDO Conc. *Nusa*, pages 77–101.

Rizka Wakhidatus Sholikah, Agus Zainal Arifin, Diana Purwitasari, and Chastine Fatichah. 2017. Co-occurrence technique and dictionary based method for Indonesian thesaurus construction. *5th International Conference on Information and Communication Technology, ICoIC7 2017*.

Deardo Dibrianto Sinaga and Seng Hansun. 2018. Indonesian text document similarity detection system using rabin-karp and confix-stripping algorithms. *International Journal of Innovative Computing, Information and Control*, 14(5):1893–1903.

Sandhya Singh, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2016. IIT Bombay's English-Indonesian submission at wat: Integrating neural language models with SMT. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 68–74.

Ahmad Hasan Siregar and Dina Chahyati. 2020. Visual question answering for monas tourism object using deep learning. *12th International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, pages 381–386.

Roswani Siregar. 2017. Teaching specific purpose translation: Utilization of bilingual contract document as parallel corpus. *English Language Teaching*, 10(7):175–182.

Samuel I. G. Situmeang, Ramosan K. Lubis, Fany J. N. Siregar, and Benyamin J. D. C. Panjaitan. 2019. Movie summarization based on Indonesian subtitles with restricted boltzmann machine. *4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, pages 338–342.

Verawaty Situmorang, Tia Elyani, Roberto Tambunan, and Yohana Gulto. 2019. Applying opinion mining technique on tourism study case: Lake Toba. *Journal of Physics: Conference Series*, 1175(1).

James Neil Sneddon. 2003. *The Indonesian Language: Its History and Role in Modern Society*. UNSW Press.

Rudy Sofyan and Bahagia Tarigan. 2018. Theme markedness in the translation of student translators. *Indonesian Journal of Applied Linguistics*, 8(1):235–243.

Shahidatul Maslina Mat Sood, Tan Kim Hua, and Bahiyah Abdul Hamid. 2020. Cyberbullying through intellect-related insults. *Jurnal Komunikasi: Malaysian Journal of Communication*, 36(1):278–297.

Dewi. Soyusiawaty and Eko. Aribowo. 2016. Designing and implementing parsing for ambiguous sentences in Indonesian language. *Journal of Theoretical and Applied Information Technology*, 84(3):339–347.

Ivan Stefanus, R.S. Joko Sarwono, and Miranti Indar Mandasari. 2017. Gmm based automatic speaker verification system development for forensics in Bahasa Indonesia. *5th International Conference on Instrumentation, Control, and Automation, ICA 2017*, pages 56–61.

Mary Fatimah Subet and Mohd Ridzuan Md Nasir. 2019. Inquisitive semantic analysis of Malay language proverbs. *Malaysian Journal of Learning and Instruction*, 16(2):227–253.

Syamsul Zahri Subir. 2019. Beyond the closet? The trends and visibility of homosexuality coverage in Malaysian newspapers, 1998 - 2012. *e-BANGI*, 16(9):13–30.

Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahma Hijazi, Rayner Alfred, and Frans Coenen. 2019. Modified framework for sarcasm detection and classification in sentiment analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(3):1175–1183.

Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, Rayner Alfred, and Frans Coenen. 2017. Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts. *8th International Conference on Information Technology, ICIT 2017*, pages 703–709.

Totok Suhardijanto, Rahmad Mahendra, Zahroh Nuriah, and Adi Budiwiyanto. 2020. The framework of multiword expression in Indonesian language. Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, pages 582–588. Association for Computational Linguistics.

Totok Suhardijanto and Deodatus Perdana Putra. 2019. Acquiring extended units of meaning: The role of learner corpus in teaching Indonesian as a foreign language. In *KEBIPAAN 2019: Proceedings of the 2nd Konferensi BIPA Tahunan by Postgraduate Program of Javanese Literature and Language Education in Collaboration with Association of Indonesian Language and Literature Lecturers, KEBIPAAN, 9 November, 2019, Surakarta, C*, page 8.

Derwin Suhartono, Aryo Pradipta Gema, Suhendro Winton, Theodorus David, Mohamad Ivan Fanany, and Aniati Murni Arymurthy. 2020. Argument annotation and analysis using deep learning with attention mechanism in Bahasa Indonesia. *Journal of Big Data*, 7(1).

Adang Suhendra, Juwita Winadwiastuti, Astie Darmayantie, and Nuke Farida. 2018. Terrorism domain corpus building using latent dirichlet allocation (LDA) and its ontology relationship building using global similarity hierarchy learning(GSHL). *11th International Conference on Information and Communication Technology and System, ICTS 2017*, 2018-January:253–257.

Gilang Julian Suherik and Ayu Purwarianti. 2017. Experiments on coreference resolution for Indonesian language with lexical and shallow syntactic features. *5th International Conference on Information and Communication Technology, ICoIC7 2017*.

Herry Sujaini. 2018. Peningkatan akurasi penerjemah bahasa daerah dengan optimasi korpus paralel. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 7(1):7–12.

Sahrul Sukardi, Meredita Susanty, Ade Irawan, and Randi Fermana Putra. 2020. Low complexity named-entity recognition for Indonesian language using bilstm-cnns. *3rd International Conference on Information and Communications Technology, ICOIACT 2020*, pages 137–142.

I Made Sukarsa, I Ketut Gede Darma Putra, Nyoman Putra Sastra, and Lie Jasa. 2018. A new framework for information system development on instant messaging for low cost solution. *Telkomnika (Telecommunication Computing Electronics and Control)*, 16(6):2799–2808.

O. R. Sulaeman, W. Gata, E. Wahyudi, M. J. Hakim, R. Subandi, R. Setiyawan, and B. Pratama. 2020. Information retrieval system to find articles and clauses in uud 1945 using vector space model method. *Journal of Physics: Conference Series*, 1471(1).

Mohamed Zain Sulaiman and Muhamad Jad Hamiza Bin Mohamad Yusoff. 2020. Bila dan mengapa 'you' menjadi 'kita': Satu analisis perbandingan Inggeris-Melayu (When and why 'you' becomes 'kita': A contrastive English-Malay analysis). *GEMA Online Journal of Language Studies*, 20(4):151–165.

S. Sulaiman, R. A. Wahid, and F. Morsidi. 2017. Feature extraction using regular expression in detecting proper noun for Malay news articles based on knn algorithm. *Journal of Fundamental and Applied Sciences*, 9(5S):210–231.

Fazal Mohamed Mohamed Sultan and Syafika Atika Binti Othman. 2021. Frasa topik dan fokus dalam Bahasa Melayu: Analisis program minimalis (Topic and focus phrase in Malay language: Minimalist program analysis). *GEMA Online Journal of Language Studies*, 21(2):195–214.

Meng Sun, Marie Stephen Leo, Eram Munawwar, Paul C Condylis, Sheng-yi Kong, Seong Per Lee, Albert Hidayat, and Muhamad Danang Kerianto. 2020. Semi-supervised category-specific review tagging on Indonesian e-commerce product reviews. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 59–63. Association for Computational Linguistics.

Dedy Suryadi. 2021. Does it make you sad? a lexicon-based sentiment analysis on covid-19 news tweets. *IOP Conference Series. Materials Science and Engineering*, 1077(1).

Endang Suryawati, Munandar Devi, Dianadewi Riswantini, Achmad Fatchuttamam Abka, and Andria Arisal. 2018. Pos-tagging for informal language (study in Indonesian tweets). *International Conference on Data and Information Science 2017, ICoDIS 2017*, 971.

Tata Sutabri and Miftah Ardiansyah. 2017. Framework of sentiment annotation for document specification in Indonesian language base on topic modeling and machine learning. *5th International Conference on Cyber and IT Service Management, CITSM 2017*.

Taufic Leonardo Sutejo and Dessi Puji Lestari. 2019. Indonesia hate speech detection using deep learning. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 39–43.

Wiwin Suwarningsih and Nuryani. 2019. Opinion qa-pairs generation from Indonesian Twitter. *6th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2018*, pages 209–213.

Suyanto Suyanto. 2019a. Flipping onsets to enhance syllabification. *International Journal of Speech Technology*, 22(4):1031–1038.

Suyanto Suyanto. 2019b. Incorporating syllabification points into a model of grapheme-to-phoneme conversion. *International Journal of Speech Technology*, 22(2):459–470.

Suyanto Suyanto. 2020. Phonological similarity-based backoff smoothing to boost a bigram syllable boundary detection. *International Journal of Speech Technology*, 23(1):191–204.

Suyanto Suyanto, Anditya Arifianto, Anis Sirwan, and Angga P. Rizaendra. 2020. End-to-end speech recognition models for a low-resourced Indonesian language. *8th International Conference on Information and Communication Technology, ICoICT 2020*.

Suyanto Suyanto, Sri Hartati, Agus Harjoko, and Dirk Van Compernolle. 2016. Indonesian syllabification using a pseudo nearest neighbour rule and phonotactic knowledge. *Speech Communication*, 85:109–118.

Suyanto Suyanto, Andi Sunyoto, Rezza Nafi Ismail, Ema Rachmawati, and Warih Maharani. 2021. Stemmer and phonotactic rules to improve n-gram tagger-based Indonesian phonemicization. *Journal of King Saud University - Computer and Information Sciences*.

Arida Ferti Syafiandini, Hani Febri Mustika, Lindung Parningotan Manik, Yan Rianto, and Zaenal Akbar. 2019. Implementing graph based rank on online news media keyword extraction. *7th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2019*, pages 108–113.

Indira Syawanodya and Arief Fatchul Huda. 2018. Improvement on stemmer algorithm for Indonesian language with spellchecker. *3rd International Conference on Informatics and Computing, ICIC 2018*.

Kathleen Swee Neo Tan, Tong Ming Lim, and Yee Mei Lim. 2020. Emotion analysis using self-training on Malaysian code-mixed Twitter data. In *13th IADIS International Conferences ICT, Society, and Human Beings 2020; Connected Smart Cities 2020; and Web Based Communities and Social Media 2020*, pages 181–188. IADIS.

Kim Hua Tan, Abdullah Imran Ho, Nur Azureen Zulkifli, and Shukor Shamirr Muhammad Mohd. 2017a. Trend penggunaan bahasa samar dalam persidangan parlimen Malaysia (Trend of adjunctive and disjunctive extenders usage in the Malaysian parliament). *GEMA Online Journal of Language Studies*, 17(4):84–100.

Tien-Ping Tan, Bali Ranaivo-Malançon, Laurent Besacier, Yin-Lai Yeong, Keng Hoon Gan, and Enya Kong Tang. 2017b. Evaluating lstm networks, hmm and wfst in Malay part-of-speech tagging. *Journal of Telecommunication, Electronic and Computer Engineering*, 9(2-9):79–83.

Yi-Fei Tan, Hai-Shuan Lam, Asyraf Azlan, and Wooi King Soo. 2016. Sentiment analysis for telco popularity on Twitter big data using a novel Malaysian dictionary. *7th International Conference on Applications of Digital Information and Web Technologies, ICADIWT 2016*, 282:112–125.

Theo Tanadi. 2018. Time series neural network model for part-of-speech tagging Indonesian language. In *International Conference on Information Technology and Digital Applications ICITDA 2017*, volume 325 of *IOP Conference Series. Materials Science and Engineering*. Institute of Physics Publishing.

Vincentius Gabriel Tandra, Yowen Yowen, Ravel Tanjaya, William Lucianto Santoso, and Nunung Nurul Qomariyah. 2021. Short message service filtering with natural language processing in Indonesian language. In *8th International Conference on ICT for Smart Society, ICISS 2021*. Institute of Electrical and Electronics Engineers Inc.

Dewa Ayu Nadia Taradhita and I Ketut Gede Darma Putra. 2021. Hate speech classification in Indonesian language tweets by using convolutional neural network. *Journal of ICT Research and Applications*, 14(3):225–239.

Natanael Taufik, Alfan Farizki Wicaksono, and Mirna Adriani. 2017. Named entity recognition on Indonesian microblog messages. *20th International Conference on Asian Language Processing, IALP 2016*, pages 358–361.

Syafi Muhammad Tauhid and Yova Ruldeviyani. 2020. Sentiment analysis of Indonesians response to influencer in social media. *7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020*, pages 90–95.

B. Tawaqal and S. Suyanto. 2021. Recognizing five major dialects in Indonesia based on mfcc and drnn. *Journal of Physics: Conference Series*, 1844(1).

Mohammad Teduh Uliniansyah, Hammam Riza, Agung Santosa, Gunarso, Made Gunawan, and Elvira Nurfadhilah. 2018. Development of text and speech corpus for an Indonesian speech-to-speech translation system. *20th Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, O-COCOSDA 2017*, 2018-January:53–57.

H. Thamrin, G. Ariyanto, E. W. Pamungkas, and Y. Sulistyono. 2018. User participation in building language repository: The case of Google Translate. *IOP Conference Series. Materials Science and Engineering*, 403(1).

Husni Thamrin, Gunawan Ariyanto, Irma Yuliana, and Wawan Joko Pranoto. 2019a. Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia. *Telkomnika (Telecommunication Computing Electronics and Control)*, 17(5):2321–2326.

Husni Thamrin, Gunawan Ariyanto, Irma Yuliana, and Dian Purworini. 2019b. An application that invites users to participate in developing repository of Bahasa Indonesia. *6th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2018*, pages 72–76.

Agustinus Theodorus, Tio Kristian Prasetyo, Reynaldi Hartono, and Derwin Suhartono. 2021. Short message service (sms) spam filtering using machine learning in Bahasa Indonesia. *3rd East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, pages 199–202.

Moch. Fadli Shadiqin Thirafi and Faisal Rahutomo. 2018. Implementation of naïve bayes classifier algorithm to categorize Indonesian song lyrics based on age. *3rd International Conference on Sustainable Information Engineering and Technology, SIET 2018*, pages 106–109.

C. Tho, Y. Heryadi, L. Lukas, and A. Wibowo. 2021. Code-mixed sentiment analysis of Indonesian language and Javanese language using lexicon based approach. *Journal of Physics: Conference Series*, 1869(1).

Cuk Tho, Arden S. Setiawan, and Andry Chowanda. 2018. Forming of dyadic conversation dataset for Bahasa Indonesia. *3rd International Conference on Computer Science and Computational Intelligence, ICCSCI 2018*, 135:315–322.

Ye Kyaw Thu, Win Pa Pa, Masao. Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). *10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 1574–1578.

Su-Hie Ting, Kee-Man Chuah, Collin Jerome, and Audea Johnson. 2021. Spotlight on LGBT in Malaysian online newspapers: Insights from textual analytics. *EDPACS*.

Su-Hie Ting, David Chen-On Then, and Oliver Guan-Bee Ong. 2020. Prestige of products and code-switching in retail encounters. *International Journal of Multilingualism*, 17(2):215–231.

Sabrina Tiun and Liew Siaw Hong. 2020. Identification of features in predicting prominent Malay words using decision tree. *Malaysian Journal of Computer Science*, 33(4):298–305.

Sabrina Tiun, Nor Fariza Mohd Nor, Azhar Jalaludin, and Anis Nadiah Che Abdul Rahman. 2020a. Word embedding for small and domain-specific Malay corpus. *6th International Conference on Computational Science and Technology, ICCST 2019*, 603:435–443.

Sabrina Tiun, Saidah Saad, Nor Fariza Mohd Nor, Azhar Jalaludin, and Anis Nadiah Che Abdul Rahman. 2020b. Quantifying semantic shift visually on a Malay domain-specific corpus using temporal word embedding approach. *Asia-Pacific Journal of Information Technology and Multimedia*, 9(2):1–10.

Parmonangan R. Togatorop, Roso Siagian, Yolanda Nainggolan, and Kaleb Simanungkalit. 2020. Implementation of ontology-based on word2vec and dbscan for part-of-speech. *5th International Conference on Sustainable Information Engineering and Technology, SIET 2020*, pages 51–56.

Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7):467–473. PMID: 30178033.

Hai-Long Trieu and Le-Minh Nguyen. 2018. Enhancing pivot translation using grammatical and morphological information. *15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017*, 781:137–151.

Hai-Long Trieu, Duc-Vu Tran, Ashwin Ittoo, and Le-Minh Nguyen. 2019. Leveraging additional resources for improving statistical machine translation on Asian low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(3):1–22.

I Nyoman Prayana Trisna, Aina Musdholifah, and Yunita Sari. 2020. Utilizing morphological features for part-of-speech tagging of Bahasa Indonesia in bidirectional lstm. *6th International Conference on Science in Information Technology, ICSITech 2020*, pages 51–56.

I Nyoman Prayana Trisna and Arif Nurwidyantoro. 2020. Single document keywords extraction in Bahasa Indonesia using phrase chunking. *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(4):1917–1925.

Sulis Triyono, Wening Sahayu, and Margana. 2020. Form and function of negation in German and Indonesian: Searching for equivalent construction of meaning. *Indonesian Journal of Applied Linguistics*, 9(3):675–684.

Tatiana Tsygankova, Francesca Marini, Stephen Mayhew, and Dan Roth. 2021. Building low-resource ner models using non-speaker annotations. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 62–69.

Ajeng Aulia Turdjai and Kusprasapta Mutijarsa. 2017. Simulation of marketplace customer satisfaction analysis based on machine learning algorithms. *2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, pages 157–162.

Mohammad Teduh Uliniansyah, Gunarso, Elvira Nurfadhilah, Lyla Ruslana Aini, Juliati Junde, Fara Ayuningtyas, and Agung Santosa. 2016. A tool to solve sentence segmentation problem on preparing speech database for Indonesian text-to-speech system. *5th*

Workshop on Spoken Language Technologies for Under-resourced languages, SLTU 2016, 81:188–193.

Mohammad Teduh Uliniansyah, Elvira Nurfadhilah, Harnum Annisa, Made Gunawan, Lyla Ruslana Aini, Agung Santosa, Asril Jarin, Gunarso, Fara Ayuningtyas, and Hammam Riza. 2019. Utilizing Indonesian allophones and intraword short pauses handling to improve performance of Indonesian text-to-speech. *22nd International Conference on Asian Language Processing, IALP 2018*, pages 143–146.

Priva Uriel Cohen. 2017. Informativity and the actuation of lenition. *Language*, 93(3):569–597.

Ema Utami, Anggit Dwi Hartanto, Sumarni Adi, Irwan Oyong, and Suwanto Raharjo. 2019a. Profiling analysis of disc personality traits based on Twitter posts in Bahasa Indonesia. *Journal of King Saud University - Computer and Information Sciences*.

Ema Utami, Anggit Dwi Hartanto, Sumarni Adi, Rahardyan Bisma. Setya Putra, and Suwanto Raharjo. 2019b. Formal and non-formal Indonesian word usage frequency in Twitter profile using non-formal affix rule. *1st International Conference on Cybernetics and Intelligent System, ICORIS 2019*, pages 173–176.

Ema Utami, Irwan Oyong, Suwanto Raharjo, Anggit Dwi Hartanto, and Sumarni Adi. 2021. Supervised learning and resampling techniques on disc personality classification using Twitter information in Bahasa Indonesia. *Applied Computing and Informatics*.

Dominique Vervoort and Jessica G. Luc. 2020. Hashtag global surgery: The role of social media in advancing the field of global surgery. *Cureus*, 12(6):e8468.

C. B. Vista, C. H. Satriawan, D. P. Lestari, and D. H. Widyantoro. 2018. Specific acoustic models for spontaneous and dictated style in Indonesian speech recognition. *2nd International Conference on Computing and Applied Informatics 2017, ICCAI 2017*, 978.

Agung Wahana, Diena Rauda Ramdania, Dhanis Al Ghifari, Ichsan Taufik, Faiz M. Kaffah, and Yana Aditia Gerhana. 2020. Breakdown film script using parsing algorithm. *Telkomnika (Telecommunication Computing Electronics and Control)*, 18(4):1976–1982.

Ummi Nadjwa Binti Wahiyudin and Taj Rijal Bin Muhamad Romli. 2021. Translating Malay compounds into Arabic based on dynamic theory and arabization method. *Journal of Islamic Thought and Civilization*, 11(1):43–58.

Eka Dyar Wahyuni, Amalia Anjani Arifiyanti, and Mohamad Irwan Afandi. 2020. School from home situation in Indonesia: An exploratory data analysis of Indonesian tweet data. *6th Information Technology International Seminar, ITIS 2020*, pages 103–108.

Lei Wang, Rong Tong, Cheung-Chi Leung, Sunil Sivadas, Chongjia Ni, and Bin Ma. 2018. Cloud-based automatic speech recognition systems for southeast Asian languages. *5th International Conference on Orange Technologies, ICOT 2017*, 2018-January:147–150.

Lianxi Wang, Xiaotian Lin, and Nankai Lin. 2021. Research on pseudo-label technology for multi-label news classification. In J. Llados, D. Lopresti, and S. Uchida, editors, *16th International Conference on Document Analysis and Recognition, ICDAR 2021*, volume 12822 LNCS, pages 683–698. Springer Science and Business Media Deutschland GmbH.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*, 42(2):277–306.

Vivien Arief Wardhany, Muhammad Hendrick Kurnia, Sritrusta Sukaridhoto, Amang Sudarsono, and Dadet Pramadihanto. 2016. Smart presentation system using hand gestures and Indonesian speech command. *17th International Electronics Symposium, IES 2015*, pages 68–72.

Harco Leslie Hendric Spits Warnars, Jessica Aurellia, and Kendrick Saputra. 2021. Translation learning tool for local language to Bahasa Indonesia using knuth-morris-pratt algorithm. *TEM Journal*, 10(1):55–62.

Tifani Warnita and Dessi Puji Lestari. 2017. Identifying deception in Indonesian transcribed interviews through lexical-based approach. *31st Pacific Asia Conference on Language, Information and Computation, PACLIC 2017*, pages 148–154.

Kok Weiying, Duc Nghia Pham, Ngo Chuan Hai, and Hong Hoe Ong. 2018. Topic modelling for Malay news aggregator. *4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018*.

Annabelle Wenas, Smita Sjahputri, Takwin Bagus, Alfindra Primaldhi, and Muhamad Roby. 2016. Measuring happiness in large population. *IOP Conference Series. Earth and Environmental Science*, 31(1).

Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. 2020. Semi-supervised low-resource style transfer of Indonesian informal to formal language with iterative forward-translation. *2020 International Conference on Asian Language Processing, IALP 2020*, pages 310–315.

Nelly Indriani Widiastuti and Maulvi Inayat Ali. 2021. Elman recurrent neural network for aspect based sentiment analysis. *Journal of Engineering Science and Technology*, 16(3):1991–2000.

W. Widodo, M. Nugraheni, and I. P. Sari. 2021. A comparative review of extractive text summarization in Indonesian language. *IOP Conference Series. Materials Science and Engineering*, 1098(3).

Bambang Dwi Wijanarko, Yaya Heryadi, Hapnes Toba, and Widodo Budiharto. 2020. Automated question generating method based on derived keyphrase structures from bloom's taxonomy. *ICIC Express Letters*, 14(11):1059–1067.

Bambang Dwi Wijanarko, Yaya Heryadi, Hapnes Toba, and Widodo Budiharto. 2021. Question generation model based on key-phrase, context-free grammar, and bloom's taxonomy. *Education and Information Technologies*, 26(2):2207–2223.

Wilbert Wijaya, I Made Murwantara, and Aditya Rama Mitra. 2020. A simplified method to identify the sarcastic elements of Bahasa Indonesia in YouTube comments. *8th International Conference on Information and Communication Technology, ICoICT 2020*.

Rini Wijayanti, Masayu Leylia Khodra, and Dwi Hendratmo Widyantoro. 2021. Indonesian abstractive summarization using pre-trained model. *3rd East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, pages 79–84.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 843–857. Association for Computational Linguistics.

Gabriella Putri Wiratama and Andre Rusli. 2019. Sentiment analysis of application user feedback in Bahasa Indonesia using multinomial naive bayes. *5th International Conference on New Media Studies, CONMEDIA 2019*, pages 223–227.

Tri Wiratno and Hisham Dzakiria. 2016. Examining the writing genre in journal articles of natural science and social science. *Advanced Science Letters*, 22(12):4431–4435.

Maulana Wisnu Prabowo, Indra Budi, and Harry Budi Santoso. 2021. Developing question generation system for Bahasa Indonesia using Indonesian standard language regulation. In *10th International Conference on Software and Computer Applications, ICSCA 2021*, pages 258–261. Association for Computing Machinery.

Sri Ratna Wulan and Suhono Harso Supangkat. 2018. Semi-supervised learning self-training for Indonesian motivational messages classification. *2017 International Conference on ICT for Smart Society, ICISS 2017*, 2018-January:1–7.

Benjamin Chu Min Xian, Mohammad Arshi Saloot, Amiera Syazreen Mohd Ghazali, Khalil Bouzekri, Rohana Mahmud, and Dickson Lukose. 2016. Benchmarking Mi-AR: Malay anaphora resolution. *2016 International Conference on Optoelectronics and Image Processing, ICOIP 2016*, pages 59–69.

Fuad Yahaya, Nurazzah Abdul Rahman, and Zainab Abu Bakar. 2017a. Resolving Malay word sense disambiguation utilizing cross-language learning sources approach. *Advanced Science Letters*, 23(11):11320–11324.

M. F. Yahaya, N. A. Rahman, Z. A. Bakar, and H. Hasmy. 2017b. Evaluation on knowledge extraction and machine learning in resolving Malay word ambiguity. *Journal of Fundamental and Applied Sciences*, 9(5S):115–130.

Mohd Fuad Yahaya, Nurazzah Abd Rahman, and Zainab Abu Bakar. 2018. Morphological analysis of Malay words for resolving ambiguity. *4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences, CAMP 2018*, pages 31–35.

Nurenzia Yannuar, Emalia Iragiliati, and Zen Evynurul Laily. 2017. Bòsò walikan malang's address practices. *GEMA Online Journal of Language Studies*, 17(1):107–123.

Muhamad Rizky Yanuar and Shun Shiramatsu. 2020. Aspect extraction for tourist spot review in Indonesian language using bert. *2nd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020*, pages 298–302.

Yangfan Yao, Jian Cao, Tao Wang, Rui Liu, Zhenyuan Dai, and Haoqiang Yuan. 2020. Efficient implementation of dirty words detection in decision tree model. *5th IEEE International Conference on Signal and Image Processing, ICSIP 2020*, pages 60–64.

Yin-Lai Yeong, Tien-Ping Tan, Keng Hoon Gan, and Siti Khaotijah Mohammad. 2018. Hybrid machine translation with multi-source encoder-decoder long short-term memory in English-Malay translation. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2):1446–1452.

Yin-Lai Yeong, Tien-Ping Tan, and Siti Khaotijah Mohammad. 2016. Using dictionary and lemmatizer to improve low resource English-Malay statistical machine translation system. *5th Workshop on Spoken Language Technologies for Under-resourced languages, SLTU 2016*, 81:243–249.

Ong Jun Ying, Muhammad Mun'im Ahmad Zabidi, Norhafizah Ramli, and Usman Ullah Sheikh. 2020. Sentiment analysis of informal Malay tweets with deep learning. *IAES International Journal of Artificial Intelligence*, 9(2):212–220.

Emny Harna Yossy, Syaeful Karim, and Dannys Dian Rachmantyo. 2020. Development of Indonesian language speech recognition algorithm model in knowledge database. *7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020*, pages 126–129.

Julio Christian Young and Andre Rusli. 2019. Review and visualization of Facebook's fasttext pretrained word vector model. *2019 International Conference on Engineering, Science, and Industrial Applications, ICESI 2019*.

Grelly Lucia Yovellia Londo, Dwiky Hutomo Kartawijaya, Hesti Tri Ivariyani, Yohanes Sigit Purnomo W.P, Aryasuta P. Muhammad Rafi, and Dipo Ariyandi. 2019. A study of text classification for Indonesian news article. *1st International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pages 205–208.

Ranu Yulianto and Siti Mariyah. 2017. Building automatic mind map generator for natural disaster news in Bahasa Indonesia. *4th International Conference on Information Technology Systems and Innovation, ICITSI 2017*, 2018-January:177–182.

Rajif Agung Yunmar and I. Wayan Wiprayoga Wisesa. 2019. Design of ontology-based question answering system for incompleted sentence problem. *International Conference on Science, Infrastructure Technology and Regional Development 2018, ICoSITeR 2018*, 258.

Ahmad Rizal Mohd Yusof, Mohd Firdaus Hamdan, Shamsul Amri Baharuddin, and Mohd Syariefudin Abdullah. 2017. Bahasa Melayu sebagai bahasa ilmu (BMBI) di ruang siber: Suatu analisis sosiologi tehadap pembangunan pangkalan data BMBI. *e-BANGI*, 12(3):1–15.

Maslida Yusof and Karim Harun. 2021. Spesifikasi ruang dalam kata kerja deiktik datang dan pergi (Spatial specification in deictic verbs datang and pergi). *Geografia*, 17(2).

Maslida Yusof, Karim Harun, and Nasrun Alias. 2016. 'Sampai Di' vs 'Sampai Ke': Accomplishment or achievement verb? *Pertanika Journal of Social Sciences and Humanities*, 24(March):49–64.

Yusmeera Yusof, Siti Zamaratol-Mai Sarah Mukari Mukari, Kartini Ahmad, Mariam Adawiah Dzulkifli, Kalaivani Chellapan, Nashrah Maamor, and Wan Syafira Ishak. 2019. Development of Malay word materials for auditory-cognitive training for older adults. *International Journal on Disability and Human Development*, 18(2):153–160.

Yusra, Muhammad Fikry, Bambang Riyanto Trilaksono, Rado Yendra, and Ahmad Fudholi. 2017. Music interest classification of Twitter users using support vector machine. *Journal of Theoretical and Applied Information Technology*, 95(11):2352–2358.

Kamal Yusuf and Dewi Puspita. 2020. Diachronic corpora as a tool for tracing etymological information of Indonesian-Malay lexicon. *Register Journal*, 13(1):153–182.

Nuhu Yusuf, Mohd Amin Mohd Yunus, Norfaradilla Wahid, and Mohd Najib Mohd Salleh. 2021. A statistical linguistic terms interrelationship approach to query expansion based on terms selection value. *3rd International Conference on Information and Communication Technology and Applications, ICTA 2020*, 1350:234–244.

Raden Sandra Yuwana, Endang Suryawati, and Hilman F. Pardede. 2019. On empirical evaluation of deep architectures for Indonesian pos tagging problem. *6th International Conference on Computer, Control, Informatics and its Applications, IC3INA 2018*, pages 204–208.

Raden Sandra Yuwana, Asri Rizki. Yuliani, and Hilman F. Pardede. 2018. On part of speech tagger for Indonesian language. *2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, 2018-January:369–372.

Nur Imanina Zabha, Zakiah Ayop, Syarulnaziah Anawar, Erman Hamid, and Zaheera Zainal Abidin. 2019. Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. *International Journal of Advanced Computer Science and Applications*, 10(1):346–351.

Muhamad Fadzllah Zaini, Anida Saruddin, Mazura Mastura Muhammad, and Siti Saniah Abu Bakar. 2020a. Perception and metaphorical smell: A Malay manuscript study (petua membina rumah) as an Asian text. In *2020 Conference of the Global Council on Anthropological Linguistics in Asia, GLOCAL 2020*, volume 2020-January, pages 345–351. Global Council on Anthropological Linguistics.

Muhamad Fadzllah Zaini, Anida Sarudin, Mazura Mastura Muhammad, and Siti Saniah Abu Bakar. 2020b. Representatif leksikal ukuran sebagai metafora linguistik berdasarkan teks klasik Melayu (Representatives of lexical ukuran as linguistics metaphors based on Malay classic text). *GEMA Online Journal of Language Studies*, 20(2):168–187.

Muhamad Fadzllah Zaini, Anida Sarudin, Mazura Mastura Muhammad, Zulkifli Osman, Husna Faredza Mohamed Redzwan, and Muhammad Aanas Al-Muhsin. 2021. House building tips (HBT) corpus dataset as a resource to discover Malay architectural ingenuity and identity. *Data in Brief*, 36:107013.

Zamahsyari and Arif Nurwidyantoro. 2017. Sentiment analysis of economic news in Bahasa Indonesia using majority vote classifier. *3rd International Conference on Data and Software Engineering, ICoDSE 2016*.

Bakar Zamri Abu, Ismail Normaly Kamal, and Rawi Mohd Izani Mohamed. 2017. Rule-based approach on extraction of Malay compound nouns in standard Malay document. *IOP Conference Series. Materials Science and Engineering*, 226(1).

Subhan Zein. 2020. *Language Policy in Superdiverse Indonesia*. Taylor & Francis Group.

Zhiping Zeng, Van Tung Pham, Haihua Xu, Yerbolat Khassanov, Eng Siong Chng, Chongjia Ni, and Bin Ma. 2021. Leveraging text data using hybrid transformer-lstm based end-to-end ASR in transfer learning. *12th International Symposium on Chinese Spoken Language Processing, ISCSLP 2021*.

Lixuan Zhao, Jian Yang, and Qinglai Qin. 2020. Enhancing prosodic features by adopting pre-trained language model in Bahasa Indonesia speech synthesis. *3rd International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2020*.

Nurul Syeilla Syazhween Zulkefli, Nurazzah Abdul Rahman, and Mazidah Puteh. 2017. A survey: Framework of an information retrieval for Malay translated hadith document. *8th International Conference on Mechanical and Manufacturing Engineering, ICME 2017*, 135.

A Appendix

A more accessible table will be available here.

Table 1

Education Specific Studies		
Title	Author	Year
The Development of an Audible Pattani Malay- . . .	Boonkwan et al.	2016
A corpus platform of Indonesian academic language	Kwary	2019
U-tapis: Automatic spelling filter as an effort to . . .	Mediyawati et al.	2021
Word level auto-correction for latent semantic . . .	Ratna et al.	2017
The development of Indonesian POS tagging sys . . .	Muljono et al.	2017c
A morphophonemic analysis on the affixation in . . .	Ampa et al.	2019
Learning Indonesian Frequently Used Vocabulary . . .	Lin et al.	2019b
Towards developing colloquial Indonesian lan . . .	Nataprawira and Carey	2020
Pengajaran bahasa dan pemerolehan bahasa ke . . .	Rosiyana	2020
Cross-corpus native language identification via . . .	Rangel et al.	2018
Acquiring Extended Units of Meaning: The Role . . .	Suhardijanto and Putra	2019
Designing Phonetic Alphabet for Bahasa Indone . . .	Karlina et al.	2020
Indonesian essay grading module using Natural . . .	Ajitiono and Widyani	2017
Automated Bahasa Indonesia essay evaluation . . .	Amalia et al.	2019a
Exploiting Syntactic Similarities for Preposition . . .	Irmawati et al.	2016
Vocabulary Load on Two Mainstream Indonesian . . .	Destiani et al.	2018a
Perbandingan Deiksis pada Dua Buku Ajar: Anal . . .	Destiani et al.	2018b
Pengembangan kamus pemelajar Bahasa Indone . . .	Fadly	2018
Teaching Specific Purpose Translation: Utiliza . . .	Siregar	2017
Theme markedness in the translation of student . . .	Sofyan and Tarigan	2018
Strategi Pengukuran Upaya Berbahasa Menerusi . . .	Redzwan et al.	2020
Generating artificial error data for Indonesian . . .	Irmawati et al.	2017b
Menangani Kekaburan Kemahiran Prosedur dan . . .	Anida et al.	2019
Environmental awareness content for character . . .	Rahmawati et al.	2020
Inquisitive semantic analysis of Malay language . . .	Subet and Md Nasir	2019
An experimental study of text preprocessing tech . . .	Hasanah et al.	2018
N-Gram Keyword Retrieval on Association Rule . . .	Setiawan et al.	2018
Semi-supervised learning self-training for Indone . . .	Wulan and Supangkat	2018
Evaluating rnn architectures for handling imbal . . .	Christianto et al.	2020
A comparison of supervised text classification . . .	Dhammajoti et al.	2020
Automatic Indonesia's questions classification . . .	Kusuma et al.	2016
Answer categorization method using K-means for . . .	Ratna et al.	2019
Knowing Right from Wrong: Should We Use . . .	Septiandri et al.	2020
WPS: Application for Generating Answer of . . .	Oktavia et al.	2021
Question generation model based on key-phrase, . . .	Wijanarko et al.	2021
Developing Question Generation System for Ba . . .	Wisnu Prabowo et al.	2021
Applications of natural language processing in . . .	Maxwell-Smith et al.	2020
Developing an online self-learning system of In . . .	Muljono et al.	2016
Automatic Pronunciation Generator for Indone . . .	Hoesen et al.	2019
Anita: Intelligent Humanoid Robot with Self- . . .	Andreas et al.	2019
A novel model and implementation of humanoid . . .	Budiharto et al.	2021

End of Table

B Appendix

A more accessible table will be available here.

Table 2

Broad NLP		
Title	Author	Year
IndoLEM and IndoBERT: A Benchmark Dataset ...	Koto et al.	2020b
Unstructured Malay Text Analytics Model in Crime	Mohemad et al.	2020c
A new framework for information system devel ...	Sukarsa et al.	2018
An overview of BPPT's Indonesian language re ...	Gunarso et al.	2016
Challenges and development in Malay natural lan ...	Lan and Logeswaran	2020
Statistical and Corpus Work		
Title	Author	Year
The Development of an Audible Pattani Malay- ...	Boonkwan et al.	2016
Linking the TUFS Basic Vocabulary to the Open ...	Bond et al.	2020
Hypernym-Hyponym Relation Extraction from ...	Nityasya et al.	2019
Linguistik Korpus Kuantitatif dan Kajian Seman ...	Rajeg	2020
Characteristics of Malay translated hadith corpus	Sazali et al.	2020
Syllabification Model of Indonesian Language ...	Fanani and Suyanto	2021
A corpus platform of Indonesian academic language	Kwary	2019
Examining the writing genre in journal articles of ...	Wiratno and Dzakiria	2016
An annotated news corpus of Malaysian Malay	Chung and Shih	2019
Introduction of the Asian Language Treebank	Riza et al.	2017
Information extraction: Evaluating named entity ...	Sazali et al.	2017
Comparative study on corpus development for ...	Din et al.	2017
Building the Pornography Corpus for Bahasa In ...	Gunawan et al.	2019c
Building a Malay-English code-switching subjec ...	Kasmuri and Basiron	2019
IndoNLU: Benchmark and Resources for Evaluat ...	Wilie et al.	2020
Development of a retrieval system for Al Hadith ...	Aulia et al.	2017b
An Ontological Approach towards Dialogue ...	Mohd Yunus et al.	2017
Co-occurrence technique and dictionary based ...	Sholikah et al.	2017
Penentuan Fitur bagi Pengekstrakan Tajuk Berita ...	Shahrul Azman Mohd et al.	2018a
The Development of the Malaysian Hansard Cor ...	Abdullah et al.	2021
Rancang bangun aplikasi web scraping untuk kor ...	Mitra et al.	2017
NUWT: Jawi-specific buckwalter corpus for ...	Bakar et al.	2016
Towards Computational Linguistics in Minangk ...	Koto and Koto	2020
Indonesia Language Sphere: an ecosystem for ...	Murakami	2019
Designing a collaborative process to create bilin ...	Nasution et al.	2019
Tufs asian language parallel corpus (talpco)	Nomoto et al.	2018b
Sentence segmentation and phrase strength esti ...	Hanum and Bakar	2016b
Evaluation of Energy and Duration on Malay ...	Hanum and Bakar	2016a
Prosodic breaks on Malay speech corpus: Evalua ...	Hanum et al.	2017
Demarcating and highlighting in Papuan Malay ...	Kaland and Baumann	2020
Repetition Reduction Revisited: The Prosody of ...	Kaland and Himmelmann	2020
Stress predictors in a Papuan Malay random forest	Kaland et al.	2019
Lexical analyses of the function and phonology ...	Kaland et al.	2021
Development of under-resourced Bahasa Indone ...	Cahyaningtyas and Arifianto	2018
Generative Indonesian Conversation Model using ...	Chowanda and Chowanda	2018
An evaluation of sentence selection methods on ...	Muljono et al.	2020
Indonesian Affective Word Resources Construc ...	Hulliyah et al.	2019

Continuation of Table 2

Title	Author	Year
Analysis of Indonesian sentiment text based on ...	Hulliyah et al.	2017
An integrated semi-automated framework for ...	Kaity and Balakrishnan	2020
BMBI: A Development of a Special Corpus on ...	Rumaisa et al.	2020
Sentiment analysis in Indonesian and French by ...	Shalunts et al.	2018
Preliminary Research Design on Sensor Data ...	Aulia et al.	2020
Review on the Role of Social Media for Dengue ...	Kannan et al.	2019
Twitter corpus creation: The case of a Malay ...	Saloot et al.	2016
An Application that Invites Users to Participate ...	Thamrin et al.	2019b
A publicly available Indonesian corpora for auto ...	Koto	2016
Development of multilingual social media data ...	Rumaisa et al.	2019
Bahasa Indonesia text corpus generation using ...	Amalia et al.	2019b
WatsaQ: Repository of Al Hadith in Bahasa (Case ...	Aulia et al.	2017a
Malay Online Virtual Integrated Corpus ...	Awang Abu Bakar et al.	2018
The Development of an Integrated Corpus for ...	Bakar	2020
Building Corpus in Bahasa Indonesia for Porno ...	Chandra et al.	2019
Building a web-based application for language ...	Dinakaramani and Suhardijanto	2019
Development of a Web-based Jahai-Malay Lan ...	Mohtar et al.	2021
A Review on Building Bilingual Comparable Cor ...	Nasharuddin et al.	2018
Linguistic studies using large annotated corpora: ...	Nomoto and Moeljadi	2019
Introducing the Asian language treebank (ALT)	Thu et al.	2016
Bahasa Melayu sebagai Bahasa Ilmu (BMBI) di ...	Yusof et al.	2017
A dependency annotation scheme to extract syn ...	Irmawati et al.	2017a
Domain-specific stop words in Malaysian parlia ...	Rahman et al.	2021a
Transforming semi-structured indigenous dictio ...	Ranaivo-Malançon et al.	2017
Rule-based text normalization for Malay social ...	Ariffin and Tiun	2020
Evaluating the use of word embeddings for part- ...	Abka	2017
Information Retrieval System to Find Articles and ...	Sulaeman et al.	2020
U-tapis: Automatic spelling filter as an effort to ...	Mediyawati et al.	2021
Building the Application to Identify Incorrect Cap ...	Gunawan et al.	2019a
Feature extraction using regular expression in de ...	Sulaiman et al.	2017
Incorporating Knowledge Base in Unsupervised ...	Rifin and Hamzah	2017
Movie Summarization based on Indonesian Subti ...	Situmeang et al.	2019
Semantic similarity measures for Malay-English ...	Mahadzir et al.	2018
Recognizing and normalizing temporal expres ...	Mirza	2016
Automatic Grammar Checking System for Indone ...	Rahutomo et al.	2018
Rapid Heteronym Disambiguation for Text-to- ...	Samsudin and Rahim	2019
Identification Of Features In Predicting Promi ...	Tiun and Hong	2020
An Enhancement of Malay Social Media Text ...	Bakar et al.	2019
Text Normalization Algorithm on Twitter in Com ...	Hanafiah et al.	2017
Pre-processing Tasks in Indonesian Twitter Messages	Hidayatullah and Ma'arif	2017
Proposal: A Hybrid Dictionary Modelling Ap ...	Nor Azlizawati Binti et al.	2017
Normalization of Indonesian-English code-mixed ...	Barik et al.	2019
Review and Visualization of Facebook's FastText ...	Young and Rusli	2019
Bidirectional encoder representations from trans ...	Candra et al.	2021
Categorization of Malay social media text and ...	Maskat and Rahman	2020
Exploring Edit Distance for Normalising Out-of- ...	Raja et al.	2019
An automatic construction of Malay stop words ...	Chekima and Alfred	2016
Word Sense Disambiguation in Bahasa Indonesia ...	Faisal et al.	2018
Cross-Lingual and Supervised Learning Ap ...	Mahendra et al.	2018b

Continuation of Table 2

Title	Author	Year
Enhancing Latent Semantic Analysis by Embed ...	Rahman et al.	2017
Word level auto-correction for latent semantic ...	Ratna et al.	2017
Evaluating Word Embeddings for Indone ...	Rizal and Stymne	2020
Designing and implementing parsing for ambigu ...	Soyusiawaty and Aribowo	2016
Word Embedding for Small and Domain-specific ...	Tiun et al.	2020a
Resolving Malay word sense disambiguation uti ...	Yahaya et al.	2017a
Morphological Analysis of Malay Words for Re ...	Yahaya et al.	2018
Evaluation on knowledge extraction and machine ...	Yahaya et al.	2017b
Naïve Bayes implementation into Bahasa Indone ...	Jodhinata and Hartanti	2016
Analyzing Malay Stemmer Performance Towards ...	Rodzman et al.	2018
Information Retrieval Technique for Indonesian ...	Riza et al.	2020
Stemmer and phonotactic rules to improve n-gram ...	Suyanto et al.	2021
Analysis of Stemming Influence on Indonesian ...	Hidayatullah et al.	2016
Towards stemming error reduction for Malay texts	Kassim et al.	2019
Enhanced Text Stemmer with Noisy Text Normal ...	Kassim et al.	2020b
Design Consideration of Malay Text Stemmer ...	Kassim et al.	2020a
Malay word stemmer to stem standard and slang ...	Kassim et al.	2016b
Enhanced rules application order to stem affixa ...	Kassim et al.	2016a
Word stemming challenges in Malay texts: A lit ...	Kassim et al.	2016c
Non-formal affixed word stemming in Indonesian ...	Putra et al.	2018a
Accuracy measurement on Indonesian non-formal ...	Putra et al.	2019
Improving stemming techniques for non-formal ...	Rianto et al.	2021
Comparison of stemming algorithms on Indone ...	Rizki et al.	2019
Improvement on stemmer algorithm for Indone ...	Syawanodya and Huda	2018
The development of Indonesian POS tagging sys ...	Muljono et al.	2017c
Semantic Role Labeling in Conversational Chat ...	Rachman et al.	2018a
POS-Tagging for informal language (study in In ...	Suryawati et al.	2018
Part-of-speech tagger for Malay social media texts	Ariffin and Tiun	2018
A comparison of different part-of-speech tagging ...	Amrullah et al.	2017
Indonesian part of speech tagging using hidden ...	Cahyani and Vindiyanto	2019
Part-of-speech (pos) tagger for Malay language ...	Gaber et al.	2020
Part of Speech Tagging for Indonesian Language ...	Handrata et al.	2019
Evaluating the Morphological and Capitalization ...	Manik et al.	2019
Morphology analysis for Hidden Markov Model ...	Muljono et al.	2017a
POS-tagging for non-English tweets: An auto ...	Munandar et al.	2017
An evaluation of MorphInd's morphological an ...	Prihantoro	2021
Rule-based Part of Speech Tagger for Indonesian ...	Purnamasari and Suwardi	2018
Evaluating lstm networks, hmm and wfst in ...	Tan et al.	2017b
Time Series Neural Network Model for Part-of- ...	Tanadi	2018
Implementation of ontology-based on Word2Vec ...	Togatorop et al.	2020
Utilizing Morphological Features for Part-of- ...	Trisna et al.	2020
On Empirical Evaluation of Deep Architectures ...	Yuwana et al.	2019
On part of speech tagger for Indonesian language	Yuwana et al.	2018
Identifying Sentence Structure in Bahasa Indone ...	Gunawan et al.	2019d
Breakdown film script using parsing algorithm	Wahana et al.	2020
Algorithm for simple sentence identification in ...	Anggraini et al.	2018
Indonesian Parsing using Probabilistic Context- ...	Cahyani et al.	2020
The effectiveness of bottom up technique with ...	Fairuzz Hiloh et al.	2018
Warning and Suggestion System on Syntax Tree ...	Haris et al.	2019

Continuation of Table 2

Title	Author	Year
A Finite State Machine Model to Determine Syl ...	Haryanto and Aripin	2019
Tackling the Low-resource Challenge for Canoni ...	Mager et al.	2020
Modification of Chu-Liu/Edmonds algorithm and ...	Nizami and Purwarianti	2017
Sentence boundary disambiguation for Indone ...	Putra et al.	2017
Rule based sentence segmentation of Indonesian ...	Raharjo et al.	2018
Ensemble technique utilization for Indonesian de ...	Rahman and Purwarianti	2017
How Similar is Similar: A Comparison of Bahasa ...	Basuki and Antaputra	2020a
New tools for old tasks: A new approach to the ...	Don and Knowles	2020
Identifying and Exploiting Definitions in Wordnet ...	Moeljadi and Bond	2016
A morphophonemic analysis on the affixation in ...	Ampa et al.	2019
A study of education-related Chinese words used ...	Kia and Su'Ad	2019
Learning Indonesian Frequently Used Vocabulary ...	Lin et al.	2019b
Towards developing colloquial Indonesian lan ...	Nataprawira and Carey	2020
Pengajaran bahasa dan pemerolehan bahasa ke ...	Rosiyana	2020
An identification of authentic narrator's name fea ...	Abd Rahman et al.	2016
The process of forming a more complex idiomatic ...	Ismail et al.	2021
Exploring Lexical Differences Between Indone ...	Lin et al.	2019c
A Corpus Driven Analysis of Representations ...	Nor Fariza Mohd et al.	2019
Exploiting Malay corpus on islamic issue using ...	Setik et al.	2018
English legalese translation into Indonesian	Dewi et al.	2021
A corpus-based analysis of English core modal ...	Oktavianti	2019
Comparison of Personal Pronoun between Arabic ...	Markhamah Abdul et al.	2017
Prosody analysis of Malay language storytelling ...	Ramli et al.	2016
Code-switching in Bruneian online retail transactions	Henry and Ho	2016
Comparison of the themes of Malaysian Friday ...	Aasim Asyafi'Ie bin Ahmad et al.	2017
Where is the Head Positioned in Indonesian Lan ...	Ansari and Suhardijanto	2019
Online-Dating Romance Scam in Malaysia: An ...	Azianura Hani et al.	2019
Conceptual structure representation of causative ...	Binti Yusof and Binti Rosly	2018
A new look at Pattani Malay Initial Geminates: a ...	Burroni et al.	2020
The particle pun in modem Indonesian and ...	Chambert-Loir	2019
Lagi in standard Malaysian Malay: Its meaning ...	Chung	2019
The Indonesian prefixes PE- and PEN-: A study ...	Denistia and Baayen	2019
Similar southeast asian languages: Corpus-based ...	Ding et al.	2016
The Design of Lexical Database for Indonesian ...	Gunawan and Amalia	2017
Automatic extraction of multiword expression can ...	Gunawan et al.	2017b
The Observation of Bahasa Indonesia Official ...	Gunawan et al.	2018a
Utterance-final particles in Klang Valley Malay	Hoogervorst	2018
Covid-19 dalam Korpus Peristilahan Bahasa ...	Kasdan et al.	2020
Gandaan Separa dalam Terminologi Bahasa ...	Kasdan et al.	2017
Compilation of Malay criminological terms from ...	Lee et al.	2019
Exploring Letter's Differences between Partial ...	Lin et al.	2019a
Hedging in the discussion sections of English and ...	Loi and Lim	2019
Formation of health science terminology by users ...	Mohamad et al.	2020c
Politeness in communication through local chil ...	Mohamad Nor et al.	2019
Translation and Markedness	Ni et al.	2018
Frequency of Verbs in Lifestyle Column in the ...	Oktavianti and Pramesti	2019
The influence of students' L1 and spoken English ...	Prihantoro	2016
Sketching the Semantic Change of Jahanam and ...	Puspita and Yusuf	2020
Vector Space Models and the usage patterns of ...	Rajeg et al.	2019

Continuation of Table 2

Title	Author	Year
Cross-corpus native language identification via ...	Rangel et al.	2018
Imbuhan meN- dengan Kata Nama Konkrit Unsur ...	Saad and Jalaluddin	2020
Ideational Grammatical Metaphors in Doctrinal ...	Saragih et al.	2017
Collocation analysis of variants of intensifies in ...	Sarudin et al.	2020a
Discourse functions of the two non-active voices ...	Shiohara et al.	2019
The Framework of Multiword Expression in In ...	Suhardijanto et al.	2020
Acquiring Extended Units of Meaning: The Role ...	Suhardijanto and Putra	2019
Bila dan Mengapa 'You' Menjadi 'Kita': Satu ...	Sulaiman and Bin Mohamad Yusoff	2020
Frasa Topik Dan Fokus Dalam Bahasa Melayu: ...	Sultan and Othman	2021
Prestige of products and code-switching in retail ...	Ting et al.	2020
Quantifying semantic shift visually on a Malay ...	Tiun et al.	2020b
Informativity and the actuation of lenition	Uriel Cohen	2017
Spesifikasi ruang dalam kata kerja deiktik datang ...	Yusof and Harun	2021
'Sampai Di' Vs 'Sampai Ke': Accomplishment ...	Yusof et al.	2016
Diachronic Corpora as a Tool for Tracing Etymo ...	Yusuf and Puspita	2020
Representatif Leksikal Ukuran sebagai Metafora ...	Zaini et al.	2020b
Designing Phonetic Alphabet for Bahasa Indone ...	Karlina et al.	2020
Indonesian essay grading module using Natural ...	Ajitiono and Widyani	2017
Automated Bahasa Indonesia essay evaluation ...	Amalia et al.	2019a
Exploiting Syntactic Similarities for Preposition ...	Irmawati et al.	2016
Vocabulary Load on Two Mainstream Indonesian ...	Destiani et al.	2018a
Perbandingan Deiksis pada Dua Buku Ajar: Anal ...	Destiani et al.	2018b
Pengembangan kamus pemelajar Bahasa Indone ...	Fadly	2018
Fossicking in dominant language teaching: Ja ...	Maxwell-Smith	2021
Teaching Specific Purpose Translation: Utiliza ...	Siregar	2017
Theme markedness in the translation of student ...	Sofyan and Tarigan	2018
Strategi Pengukuran Upaya Berbahasa Menerusi ...	Redzwan et al.	2020
Generating artificial error data for Indonesian ...	Irmawati et al.	2017b
Menangani Kekaburan Kemahiran Prosedur dan ...	Anida et al.	2019
Environmental awareness content for character ...	Rahmawati et al.	2020
Development of Malay word materials for ...	Yusof et al.	2019
Exploring gender issues associated with ...	Aziz	2019
"Happiness" in Bahasa Indonesia and its implica ...	Effendi and Muchammadun	2018
Defying the global: The cultural connotations of ...	Hashim and Rahim	2016
The implicit meaning in Malay figurative lan ...	Mansor and Jalaluddin	2016
"Is Selangor in Deep Water?": A Corpus-driven ...	Norsimah Mat et al.	2019
Linguistic Representation of Violence in Judicial ...	Othman et al.	2019
Text mining of online job advertisements to iden ...	Panggih Kusuma et al.	2020
Inquisitive semantic analysis of Malay language ...	Subet and Md Nasir	2019
Spotlight on LGBT in Malaysian online newspa ...	Ting et al.	2021
The polarity of war metaphors in sports news: A ...	Hua et al.	2021
Communicating insults in cyberbullying	Hua et al.	2019
Analisis korpus terhadap idiom Bahasa Indonesia ...	Paramarta	2018
Conceptual metaphor and linguistic manifesta ...	Saad et al.	2018b
The relationship between astronomy and architec ...	Sarudin et al.	2020b
Beyond the closet? The trends and visibility of ...	Subir	2019
Trend Penggunaan Bahasa Samar dalam Persidan ...	Tan et al.	2017a
Form and function of negation in German and ...	Triyono et al.	2020
Bòsò Walikan Malang's Address Practices	Yannuar et al.	2017

Continuation of Table 2

Title	Author	Year
Perception and metaphorical smell: A Malay ...	Zaini et al.	2020a
House building tips (HBT) corpus dataset as a ...	Zaini et al.	2021
Understanding quotation extraction and attribu ...	Purnomo W.P et al.	2020
An automatic health surveillance chart interpreta ...	Aulia and Barmawi	2016
A text representation model using Sequential ...	Alias et al.	2018b
Relationship analysis of keyword and chapter in ...	Chua and Nohuddin	2017
Relation extraction using dependency tree kernel ...	Esperanti and Purwarianti	2016
An experimental study of text preprocessing tech ...	Hasanah et al.	2018
Relation Detection for Indonesian Language Us ...	Hasudungan and Purwarianti	2019
Classification of short possessive clitic pronoun ...	Noor et al.	2020
Assessing Suitable Word Embedding Model for ...	Phua et al.	2020
Experiments on coreference resolution for Indone ...	Suherik and Purwarianti	2017
Malay manuscripts transliteration using statistical ...	Razak et al.	2019
Transliteration engine for union catalogue of ...	Razak et al.	2018
SMVS: A Web-based Application for Graphical ...	Ahmat Baseri et al.	2020
Exploring Multilingual Syntactic Sentence Repre ...	Liu et al.	2019a
Transfer Building of Multiword Expression Re ...	Liu and Wang	2020
Reclassification of the Leipzig Corpora Collec ...	Nomoto et al.	2018a
Learning Indonesian-Chinese Lexicon with Bilin ...	Qiu and Zhu	2016

Machine Reading

Title	Author	Year
Towards corpus and model: Hierarchical ...	Fu et al.	2021
Towards a Standardized Dataset on Indonesian ...	Khairunnisa et al.	2020
Semi-supervised learning approach for Indone ...	Aryoyudanta et al.	2017
Named entity recognition for extracting concept ...	Santoso et al.	2021
Rule-based Approach on Extraction of Malay ...	Zamri Abu et al.	2017
A review of named entity recognition and classifi ...	Mohemad et al.	2020a
Detecting proper nouns in Indonesian-language ...	Raharjo et al.	2020
Named entity recognition on Indonesian tweets ...	Azarine et al.	2019
Named entity recognition on Indonesian Twitter ...	Rachman et al.	2018b
An enhanced Malay named entity recognition us ...	Asmai et al.	2018
Named entity recognition using fuzzy c-means ...	Salleh et al.	2018
Named entity recognition on Indonesian mi ...	Taufik et al.	2017
DBpedia entities expansion in automatically build ...	Alfina et al.	2017
Entity annotation WordPress plugin using ...	Aprilius et al.	2017
Developing name entity recognition for structured ...	Azzahra et al.	2020
Detection of compound word with combination ...	Bakar et al.	2017
Identification of Noun + Verb Compound Nouns ...	Bakar et al.	2018a
Automatic detection of compound word in Malay ...	Bakar et al.	2018b
Named-Entity Recognition for Indonesian Lan ...	Gunawan et al.	2018c
A Concise Review of Named Entity Recognition ...	Ikhwan Syafiq et al.	2019
Empirical Evaluation of Character-Based Model ...	Kurniawan and Louvan	2018
A Semi-supervised Algorithm for Indonesian ...	Leonandya et al.	2016
Malay name entity recognition using limited re ...	Noor et al.	2016
Medical Named Entity Recognition for Indone ...	Rahman	2018
Pendekatan Teknik Pengecaman Entiti Nama ...	Saad and Mohamed Kamil	2018
A Malay named entity recognition using condi ...	Salleh et al.	2017
Low Complexity Named-Entity Recognition for ...	Sukardi et al.	2020
Building Low-Resource NER Models Using Non- ...	Tsygankova et al.	2021

Continuation of Table 2

Title	Author	Year
Hate speech detection in the Indonesian language: ...	Alfina et al.	2018
Developing Indonesian corpus of pornography ...	Andriansyah et al.	2018
Bahasa Indonesia pre-trained word vector genera ...	Putri et al.	2021
Indonesian text document similarity detection sys ...	Sinaga and Hansun	2018
N-Gram Keyword Retrieval on Association Rule ...	Setiawan et al.	2018
The Effectiveness of Using Malay Affixes for Han ...	Mohamed et al.	2018
Author-Topic Modelling for Reviewer Assign ...	Kusumawardani and Khairunnisa	2019
Benchmarking Mi-AR: Malay anaphora resolution	Xian et al.	2016
Fake news identification characteristics using ...	Al-Ash and Wibowo	2018
Graph-based text representation for Malay trans ...	Alias et al.	2017a
Building automatic mind map generator for natu ...	Yulianto and Mariyah	2017
Neural sequence-to-sequence learning of internal ...	Ruzsics and Samardzic	2017
Classification of user comment using word2vec ...	Kurnia and Girsang	2021
Short Message Service (SMS) Spam Filtering us ...	Theodorus et al.	2021
Analysis and implementation of cross lingual ...	Dewi et al.	2018
Long short-term memory for hate speech and abu ...	Salim and Suhartono	2021
Semi-supervised learning self-training for Indone ...	Wulan and Supangkat	2018
Multi-Label Topic Classification of Hadith of ...	Abu Bakar et al.	2019a
Hoax analyzer for Indonesian news using rnns ...	Adipradana et al.	2021
An evolutionary-based term reduction approach ...	Alfred et al.	2017
Assessing factors that influence the performances ...	Alfred et al.	2016
A comparison study of document clustering using ...	Amalia et al.	2020a
An Efficient Text Classification Using fastText ...	Amalia et al.	2020b
Optimizing Deep Learning for Detection Cyber ...	Anindyati et al.	2019
Evaluating rnn architectures for handling imbal ...	Christianto et al.	2020
A comparison of supervised text classification ...	Dhammajoti et al.	2020
Classifying Medical Document in Bahasa Indone ...	Dhomas Hatta and Kiki Purnama	2021
Using naïve bayes classifier for application feed ...	Ferdino and Rusli	2019
The identification of pornographic sentences in ...	Gunawan et al.	2019e
The Best Parameter Tuning on RNN Layers for ...	Hikmah et al.	2020
A language identifier for Indonesian and Malay ...	Indra et al.	2016
A category classification algorithm for Indonesian ...	Jaafar et al.	2016
The impacts of singular value decomposition al ...	Jambak et al.	2019
Automatic Indonesia's questions classification ...	Kusuma et al.	2016
Comparative Study of Machine Learning Ap ...	Mohammad Najib et al.	2017
Hoax Analyzer for Indonesian News Using Deep ...	Nayoga et al.	2021
Study of hoax news detection using naïve bayes ...	Pratiwi et al.	2018
Building a question classification model for a ...	Puteh et al.	2019
Age Group Based Document Classification in Ba ...	Putra et al.	2020
Hoax web detection for news in bahasa using sup ...	Rahmat et al.	2019
Indonesian news classification using convolu ...	Ramdhani et al.	2020
Answer categorization method using K-means for ...	Ratna et al.	2019
Identifying fake news in Indonesian via super ...	Rusli et al.	2020
Indonesian news classification based on NaBaNA	Septian et al.	2017
Knowing Right from Wrong: Should We Use ...	Septiandri et al.	2020
Enhancing text classification performance by pre ...	Setiabudi et al.	2021
Text Classification Services Using Naïve Bayes ...	Setiani and Ce	2018
Argument annotation and analysis using deep ...	Suhartono et al.	2020
Semi-supervised Category-specific Review Tag ...	Sun et al.	2020

Continuation of Table 2

Title	Author	Year
Short Message Service Filtering with Natural Lan ...	Tandra et al.	2021
Implementation of Naïve Bayes Classifier Algo ...	Thirafi and Rahutomo	2018
Research on Pseudo-label Technology for Multi- ...	Wang et al.	2021
Efficient Implementation of Dirty Words Detec ...	Yao et al.	2020
A Study of Text Classification for Indonesian ...	Yovellia Londo et al.	2019
Developing the COVID-19 Malay Corpus Using ...	Hakimi and Rahman	2021
Query rewriting and corpus of semantic similarity ...	Purnamasari et al.	2016
Performance Evaluation of Inverted Files, B-Tree ...	Rosnan et al.	2019
Word prediction algorithm in resolving ambiguity ...	Sazali et al.	2016
Implementation of LSI method on information ...	Pardede and Barmawi	2016
A document recommendation system of stem ...	Parwita	2020
Weighted inverse document frequency and vector ...	Pratama et al.	2020
Cross Language Information Retrieval Using Par ...	Rahmanda et al.	2019
Machine Learning Approach for Sentiment Anal ...	Mantoro et al.	2020
Natural Language Interface to Database (NLIDB) ...	Anisyah et al.	2019
A Survey on Context-Aware Information Re ...	Bin Rodzman et al.	2018a
The implementation of fuzzy logic controller for ...	Bin Rodzman et al.	2018b
Experiment with text summarization as a positive ...	Bin Rodzman et al.	2019b
Indonesian document retrieval using vector space ...	Fitriasari et al.	2017
Access to Relational Databases Using Interroga ...	Ghassani and Widagdo	2018
Automatic open domain information extraction ...	Gultom and Wibowo	2018
Open Text Ontology Mining to Improve Re ...	Hamzah and Kamaruddin	2021
Multi-word similarity and retrieval model for a ...	Hanum et al.	2019
Syntactic rule-based approach for extracting con ...	Husin et al.	2018
Web Service for Search Engine Bahasa Indonesia ...	Husni et al.	2020
Information Retrieval for Malay Text: A Decade ...	Kamaruddin et al.	2021
Teknik Pengukuhan Perangkak Tumpuan melalui ...	Masnizah et al.	2018
Natural Language Interface to Database (NLIDB) ...	Poetra et al.	2019
Content-based Filtering Model for Recommenda ...	Putri et al.	2019b
Fabricated and Shia Malay translated hadith as ...	Rodzman et al.	2020
Domain specific concept ontologies and text sum ...	Rodzman et al.	2019
Rule-based Indonesian Open Information Extraction	Romadhony et al.	2018
A Statistical Linguistic Terms Interrelationship ...	Yusuf et al.	2021
A Survey: Framework of an Information Retrieval ...	Zulkefli et al.	2017
Crowdsourcing in developing repository of phrase ...	Thamrin et al.	2019a
Single document keywords extraction in Bahasa ...	Trisna and Nurwidyantoro	2020
Topic modeling on Indonesian online shop chat	Hidayatullah et al.	2019
Indonesian abstractive text summarization using ...	Adelia et al.	2019
Topic labeling towards news document collection ...	Adhitama et al.	2017
MYTextSum: A Malay Text Summarizer Model ...	Alias et al.	2018a
A Malay text corpus analysis for sentence com ...	Alias et al.	2016
Extract, compress and summarize–An experiment ...	Alias et al.	2017c
A Malay text summarizer using pattern-growth ...	Alias et al.	2017b
Understanding Human Sentence Compression Pat ...	Alias et al.	2018c
Bilingual extractive text summarization model ...	Alias et al.	2020
A Syntactic-based Sentence Validation Technique ...	Alias et al.	2021
Indonesian Automatic Text Summarization Based ...	Cai et al.	2019
Summarizing Indonesian news articles using ...	Garmastewira and Khodra	2019
Review of the recent research on automatic text ...	Gunawan and Amalia	2018

Continuation of Table 2

Title	Author	Year
Multi-document Summarization by using Tex ...	Gunawan et al.	2019b
Automatic Text Summarization for Indonesian ...	Gunawan et al.	2017a
Liputan6: A Large-scale Indonesian Dataset for ...	Koto et al.	2020a
Peringkasan dokumen berita Bahasa Indonesia ...	Mandar and Gunawan	2017
The purpose of bellman-ford algorithm to summa ...	Maylawati et al.	2020
Sequential pattern mining and deep learning to ...	Maylawati et al.	2019
Technique on Malay text summarization: A review	Mohemad et al.	2020b
Generation of news headline for Malay language ...	Noah et al.	2018
Text simplification for Malay corpus: A Review	Omar et al.	2021
Automatic Text Summarization for Malay News ...	Rahman et al.	2021b
Towards product attributes extraction in Indone ...	Rif'at et al.	2018
Multidocument Abstractive Summarization using ...	Severina and Khodra	2019
Penjanaan Ringkasan Isi Utama Berita Bahasa ...	Shahrul Azman Mohd et al.	2018b
Terrorism domain corpus building using Latent ...	Suhendra et al.	2018
Topic Modelling for Malay News Aggregator	Weiying et al.	2018
A comparative review of extractive text summa ...	Widodo et al.	2021
Indonesian Abstractive Summarization using Pre- ...	Wijayanti et al.	2021
A Conceptual Framework for Malay-English ...	Lim et al.	2021
Design of Ontology-based Question Answering ...	Yunmar and Wayan Wiprayoga Wisesa	2019
Corpus development for Indonesian consumer- ...	Hakim et al.	2018
Towards question identification from online ...	Mahendra et al.	2018a
WPS: Application for Generating Answer of ...	Oktavia et al.	2021
Automated question generating method based on ...	Wijanarko et al.	2020
Question generation model based on key-phrase, ...	Wijanarko et al.	2021
Developing Question Generation System for Ba ...	Wisnu Prabowo et al.	2021
Developing an adaptive language model for Ba ...	Hidayatullah and Suyanto	2019
Pembangunan Taksonomi dari Teks Melayu ...	Mohd Zakree Ahmad et al.	2018
Document Similarity Detection Using Indonesian ...	Ramadhanti and Mariyah	2019
Paraphrase construction of Al Quran in Indone ...	Hutami et al.	2019
Rude-Words Detection for Indonesian Speech Us ...	Novitasari et al.	2019
Taxonomy development from Malay text using ...	Ahmad Nazri et al.	2018
Cross-Language Plagiarism Detection System Us ...	Anak Agung Putri et al.	2017
Plagiarism Detection for Indonesian Language ...	Arifin et al.	2018
Knowledge representation system for copula sen ...	Cahyani et al.	2016
Keyword extraction from scientific articles in Ba ...	Gunawan et al.	2020
Extracting disease-symptom relationships from ...	Halim et al.	2018
Segregation of Code-Switching Sentences using ...	Kasmuri and Basiron	2020
Automated verbalization of ORM models in ...	Lim and Halpin	2016
Noun phrases extraction using shallow parsing ...	Santoso et al.	2016
Implementing Graph Based Rank on Online News ...	Syafiandini et al.	2019

Translation

Title	Author	Year
A framework for English and Malay cross-lingual ...	Nasharuddin et al.	2019
User participation in building language reposi ...	Thamrin et al.	2018
Google vs. Instagram Machine Translation: Mul ...	Larassati et al.	2019
Neural Machine Translation model for University ...	Aneja et al.	2020
Quality translation enhancement using sequence ...	Ayu et al.	2018
Meaning preservation in Example-based Machine ...	Chua et al.	2017
English-Indonesian Neural Machine Translation ...	Dwiastuti	2019

Continuation of Table 2

Title	Author	Year
Benchmarking multidomain English-Indonesian ...	Guntara et al.	2020
A Neural Machine Translation Approach for ...	Low et al.	2020
IIT Bombay's English-Indonesian submission at ...	Singh et al.	2016
Source language adaptation approaches for ...	Wang et al.	2016
Hybrid machine translation with multi-source ...	Yeong et al.	2018
Using Dictionary and Lemmatizer to Improve ...	Yeong et al.	2016
Semi-Supervised Low-Resource Style Transfer ...	Wibowo et al.	2020
Morphological analysis of speech translation into ...	Nurilman Baehaqi et al.	2019
Pengaruh Pennambahan Korpus Paralel pada ...	Abidin and Permata	2021
Effect of mono corpus quantity on statistical ma ...	Abidin et al.	2021
Peningkatan Mesin Penerjemah Statistik dengan ...	Darwis et al.	2019
Peningkatan Akurasi Penerjemah Bahasa Daerah ...	Sujaini	2018
Translation Learning Tool for Local Language to ...	Warnars et al.	2021
Leveraging additional resources for improving ...	Trieu et al.	2019
Rule-based Reordering and Post-Processing for ...	Mawalim et al.	2017
A novel Hadith authentication mobile system in ...	Fadele et al.	2020
Translation of idioms from Arabic into Malay via ...	Abidin et al.	2020
Multiple pivots in statistical machine translation ...	Budiwati and Aritsugi	2019
A Parallel Evaluation Data Set of Software Docu ...	Buschbeck and Exel	2020
A Comprehensive Analysis of Bilingual Lexicon ...	Irvine and Callison-Burch	2017
Malay-corpus-enhanced Indonesian-Chinese neu ...	Liu and Wang	2019
Language Resource Extension for Indonesian- ...	Liu et al.	2019b
Development of mobile application for Malay ...	Rahman et al.	2020
Enhancing Pivot Translation Using Grammatical ...	Trieu and Nguyen	2018
Translating Malay Compounds into Arabic Based ...	Wahiyudin and Romli	2021
Generating image description on Indonesian lan ...	Nugraha et al.	2019
Visual question answering for monas tourism ob ...	Siregar and Chahyati	2020
Learning translations via images with a massively ...	Hewitt et al.	2018
Adaptive Attention Generation for Indonesian Im ...	Mahadi et al.	2020
Cross-lingual projection for class-based language ...	Gfeller et al.	2016
Extremely Low-Resource Neural Machine Trans ...	Rubino et al.	2020
Machine translation of Indonesian: A review	Septarina et al.	2019
See also - A review on Indonesian machine trans ...	Rahutomo et al.	2019

Spoken Dialogue Systems

Title	Author	Year
Malay speech corpus of telecommunication call ...	Draman et al.	2017
Detection of Malay phrase breaks using energy ...	Mohamed Hanum and Abu Bakar	2016
A hybrid approach for single channel speech en ...	Jamal et al.	2020
Developing ASR for Indonesian-English Bilin ...	Maxwell-Smith and Foley	2021
Applications of natural language processing in ...	Maxwell-Smith et al.	2020
Robust Feature Extraction Based On Spectral And ...	Ibrahim et al.	2019
Transfer learning with bottleneck feature net ...	Lim et al.	2016
Cross-Lingual Machine Speech Chain for Ja ...	Novitasari et al.	2020
Comparing statistical classifiers for emotion clas ...	Hamzah et al.	2017
Influences of age in emotion recognition of spon ...	Jamil et al.	2017
Influences of languages in speech emotion recog ...	Rajoo and Aun	2016
Voice-Based Malay Commands Recognition by ...	Abu et al.	2020
Automatic Transcription and Captioning System ...	Andra and Usagawa	2020
Improved Transcription and Speaker Identifica ...	Andra and Usagawa	2021

Continuation of Table 2

Title	Author	Year
Speech-to-Text Conversion in Indonesian Lan ...	Dwijayanti et al.	2021
Comparison of feature extraction MFCC and LPC ...	Endah et al.	2017
Development of language identification system ...	Gunawan et al.	2018b
Voiced and unvoiced separation in Malay speech ...	Hanifa et al.	2019
Wavelet based feature extraction for the vowel sound	Hidayat et al.	2016
Shared-hidden-layer deep neural network for ...	Hoesen et al.	2018
Classification and clustering to identify spoken ...	Ibrahim and Lestari	2018
Automatic phoneme identification for Malay dialects	Khaw et al.	2017
Speech to Text of Patient Complaints for Bahasa ...	Laksono et al.	2019
Malay language speech recognition for preschool ...	Maseri and Mamat	2019
Indonesian audio-visual speech corpus for multi ...	Maulana and Fanany	2018a
Sentence-level Indonesian lip reading with spa ...	Maulana and Fanany	2018b
Sphinx4 for Indonesian continuous speech recog ...	Muljono et al.	2017b
Indonesian graphemic syllabification using a near ...	Parande and Suyanto	2019
Rule-Based Pronunciation Models to Handle ...	Putri et al.	2019a
Speech to Text Translation for Malay Language	Rami Ali and Rini	2017
Assessing automatic speech recognition in mea ...	Rosdi et al.	2017
Hybrid methods of Brandt's generalised likeli ...	Seman and Norazam	2019
Incorporating syllabification points into a model ...	Suyanto	2019b
Flipping onsets to enhance syllabification	Suyanto	2019a
Phonological similarity-based backoff smoothing ...	Suyanto	2020
End-to-End Speech Recognition Models for a ...	Suyanto et al.	2020
Indonesian syllabification using a pseudo nearest ...	Suyanto et al.	2016
Recognizing Five Major Dialects in Indonesia ...	Tawaqal and Suyanto	2021
Specific acoustic models for spontaneous and dic ...	Vista et al.	2018
Cloud-based automatic speech recognition sys ...	Wang et al.	2018
Smart presentation system using hand gestures ...	Wardhany et al.	2016
Leveraging Text Data Using Hybrid Transformer- ...	Zeng et al.	2021
Indonesian Corpus Constructing and Text Process ...	Kong and Yang	2018
Utilizing Indonesian Allophones and Intraword ...	Uliniansyah et al.	2019
Developing an online self-learning system of In ...	Muljono et al.	2016
Automatic Pronunciation Generator for Indone ...	Hoesen et al.	2019
Multi Speaker Speech Synthesis System for In ...	Budiman and Lestari	2020
A Bilingual Speech Synthesis System of Standard ...	Chen et al.	2020
Poetry visualization in digital technology	Noh et al.	2019
The first Malay language storytelling text-to- ...	Ramli et al.	2017
An Iterated Two-Step Sinusoidal Pitch Contour ...	Ramli et al.	2021
A Tool to Solve Sentence Segmentation Problem ...	Uliniansyah et al.	2016
Enhancing Prosodic Features by Adopting Pre- ...	Zhao et al.	2020
Anita: Intelligent Humanoid Robot with Self- ...	Andreas et al.	2019
A novel model and implementation of humanoid ...	Budiharto et al.	2021
Teach your robot your language! trainable neural ...	Hinaut and Twiefel	2020
The Architecture of Speech-to-Speech Translator ...	Santosa et al.	2019
Development of text and speech corpus for an ...	Teduh Uliniansyah et al.	2018
Chatbot Application on Internet of Things (IoT) ...	Gunawan et al.	2019f
Virtual assistant using lstm networks in Indonesian	Mirwan et al.	2018
Forming of Dyadic Conversation Dataset for Ba ...	Tho et al.	2018
Development of Indonesian Language Speech ...	Yossy et al.	2020
GMM based automatic speaker verification sys ...	Stefanus et al.	2017

Continuation of Table 2		
Title	Author	Year
Recurrent Neural Network to Deep Learn Conver ...	Chowanda and Chowanda	2017
Virtual phone discovery for speech synthesis with ...	Nayak et al.	2019
Speaker States		
Title	Author	Year
An automatic lexicon generation for Indonesian ...	Ayu et al.	2019
Minimally-supervised sentiment lexicon induc ...	Darwich et al.	2017
Extraction Sentiment Analysis Using naive Bayes ...	Jaka Harjanta and Herlambang	2020
Automatic Semantic Orientation of Adjectives for ...	Riyanti et al.	2018
Enhanced Malay sentiment analysis with an en ...	Al-Moslmi et al.	2017
Aspect and Opinion Extraction of Indonesian Lip ...	Kun Indarta and Romadhony	2021
Identifying deception in Indonesian transcribed ...	Warnita and Lestari	2017
Aspect Extraction for Tourist Spot Review in In ...	Yanuar and Shiramatsu	2020
Framework of sentiment annotation for document ...	Sutabri and Ardiansyah	2017
Evaluation of support vector machine and deci ...	Saad et al.	2018a
Sentiment analysis for low resource languages: A ...	Le et al.	2016
Indonesian Lexicon-Based Sentiment Analysis of ...	Kurniawan et al.	2021
A Simplified Method to Identify the Sarcastic Ele ...	Wijaya et al.	2020
Experiment with lexicon based techniques on ...	Bin Rodzman et al.	2019a
Implementation of a Machine Learning Algo ...	Buntoro et al.	2021
Sentiment Analysis of Malay Social Media Text	Chekima and Alfred	2018
Random forest approach fo sentiment analysis in ...	Fauzi	2018
Word embedding comparison for Indonesian lan ...	Imaduddin et al.	2019
Text Mining and Support Vector Machine for Sen ...	Imamah et al.	2020
Unsupervised aspect-based sentiment analysis on ...	Sasmita et al.	2018
Elman recurrent neural network for aspect based ...	Widiastuti and Ali	2021
Multilingual sentiment analysis: A systematic lit ...	Abdullah and Rusli	2021
Polarity classification tool for sentiment analysis ...	Abu Bakar et al.	2019b
Long short term memory convolutional neural ...	Af'idah et al.	2020
Malay sentiment analysis based on combined clas ...	Al-Saffar et al.	2018
An analysis of Malay language emotional speech ...	Apandi and Jamil	2017
Aspect-Based Sentiment Analysis Using Convo ...	Cahyadi and Khodra	2018
Rule-Based Model for Malay Text Sentiment ...	Chekima et al.	2018
Speech-Emotion Detection in an Indonesian Movie	Fahmi et al.	2020
A comparative study of sentiment analysis using ...	Fikri and Sarno	2019
Sentiment analysis for Malay language: system ...	Handayani et al.	2018
Sentiment analysis using recurrent neural ...	Kurniasari and Setyanto	2020a
Sentiment Analysis using Recurrent Neural Network	Kurniasari and Setyanto	2020b
Aspect-based Opinion Mining on Beauty Product ...	Mahfiz and Romadhony	2020
Aspect-Based Sentiment Analysis on Candidate ...	Manik et al.	2020
Sentiment Analysis Using Word2vec and Long ...	Muhammad et al.	2021
English and Malay cross-lingual sentiment lexi ...	Nasharuddin et al.	2017
Word2vec for Indonesian sentiment analysis to ...	Nawangsari et al.	2019
The Influence of Negation Handling on Sentiment ...	Ningtyas and Herwanto	2018
Sentiment analysis system for movie review in ...	Nurdiansyah et al.	2018
An experimental study of lexicon-based sentiment ...	Pamungkas and Putri	2017
A comparison of the use of several different re ...	Pratama et al.	2019
Pair Extraction of Aspect and Implicit Opinion ...	Setiowati et al.	2019
Sentiment analysis of application user feedback ...	Wiratama and Rusli	2019
Sentiment analysis of economic news in Bahasa ...	Zamahsyari and Nurwidyantoro	2017

Continuation of Table 2

Title	Author	Year
Indonesia Hate Speech Detection Using Deep ...	Sutejo and Lestari	2019
Criminality recognition using machine learning ...	Malim et al.	2019
Personality Measurement Design for Ontology ...	Alamsyah et al.	2020
A preliminary study on hybrid sentiment model ...	Eshak et al.	2018
A Progress on the Personality Measurement ...	Alamsyah et al.	2019
Speech Emotion Recognition for Indonesian Lan ...	Lasiman and Lestari	2019

Social Media

Title	Author	Year
Opinion QA-Pairs Generation from Indonesian ...	Suwarningsih and Nuryani	2019
Preprocessing for crawler of short message social ...	Ariestya et al.	2018
Pola Penggunaan Bahasa Melayu dalam Twitter ...	Khalid and Rahim	2021
Cyberbullying through intellect-related insults	Sood et al.	2020
Formal and Non-Formal Indonesian Word Usage ...	Utami et al.	2019b
Ten-year compilation of #savekpk Twitter dataset	Rahutomo et al.	2020
Word Cloud Result of Mobile Payment User Re ...	Dewi et al.	2020
Ensemble method for Indonesian Twitter hate ...	Fauzi and Yuniarti	2018
Event detection in Twitter: A keyword volume ...	Hossny and Mitchell	2019
Multi-label Hate Speech and Abusive Language ...	Ibrohim and Budi	2019a
Identification of hate speech and abusive language ...	Ibrohim et al.	2019
Classification of Radicalism Content from Twitter ...	Idris et al.	2019
Hierarchical multi-label classification to identify ...	Prabowo et al.	2019
Topic classification and clustering on Indonesian ...	Pratama and Purwarianti	2017
Twitter Topic Modeling on Football News	Hidayatullah et al.	2018
Topic Summarization of Microblog Document in ...	Jiwanggi and Adriani	2016
Negation handling in sentiment classification us ...	Amalia et al.	2018
A Framework for Sentiment Analysis Implemen ...	Asniar and Aditya	2017
Social Media Analytics using Sentiment and Con ...	Balakrishnan et al.	2021
Multi-Classes Emotion Detection for Unbalanced ...	Farsiah et al.	2020
Twitter sentiment analysis in under-resourced lan ...	Ferdiana et al.	2019
Corpus Usage for Sentiment Analysis of a Hash ...	Herlawati et al.	2019
Sentiment analysis on Bahasa Indonesia tweets ...	Iswanto and Poerwoto	2018
Social tension and crime related events detection ...	Jamil et al.	2019
Bilingual sentiment detection - Investigating im ...	Kaur and Balakrishnan	2016
Comparison of SVM Naïve Bayes Algorithm for ...	Kristiyanti et al.	2019
Indonesian Twitter Sentiment Analysis Using ...	Kurniawan and Maharani	2020
Aspect-level Sentiment Analysis for Social Media ...	Kusumawardani and Maulidani	2020
Sentiment Analysis Using Weighted Emoticons ...	Maulidiah Elfajr and Sarno	2018
Employ Twitter data to perform sentiment analy ...	Mohamad et al.	2020b
Classification of Twitter data by sentiment analy ...	Mohamad et al.	2020a
Detecting candidates of depression, anxiety and ...	Nasrudin et al.	2019
Naive Bayes as opinion classifier to evaluate stu ...	Permana et al.	2017
Sentiment Analysis of BPJS Kesehatan's Services ...	Rasyada et al.	2020
When Homecoming is not Coming: 2021 Home ...	Sandra and Lumbangaol	2021
Aspect-Based Sentiment Analysis for Posts on ...	Setik et al.	2021
Applying Opinion Mining Technique on Tourism ...	Situmorang et al.	2019
Does it make you sad? A lexicon-based sentiment ...	Suryadi	2021
Emotion analysis using self-training on ...	Tan et al.	2020
Sentiment analysis for telco popularity on Twitter ...	Tan et al.	2016
Hate speech classification in Indonesian language ...	Taradhita and Putra	2021

Continuation of Table 2

Title	Author	Year
Sentiment Analysis of Indonesians Response to ...	Tauhid and Ruldeviyani	2020
Code-mixed sentiment analysis of Indonesian lan ...	Tho et al.	2021
Simulation of marketplace customer satisfaction ...	Turdjai and Mutijarsa	2017
Hashtag Global Surgery: The Role of Social Me ...	Vervoort and Luc	2020
School from home situation in Indonesia: An ex ...	Wahyuni et al.	2020
Measuring happiness in large population	Wenas et al.	2016
Sentiment analysis of informal Malay tweets with ...	Ying et al.	2020
Developing cross-lingual sentiment analysis of ...	Zabha et al.	2019
Automatic Labelling of Malay Cyberbullying ...	Maskat et al.	2020
Personality prediction based on Twitter informa ...	Ong et al.	2017
Personality Modelling of Indonesian Twitter ...	Ong et al.	2021
Profiling analysis of DISC personality traits based ...	Utami et al.	2019a
Supervised learning and resampling techniques ...	Utami et al.	2021
D-Loc Apps: A Location Detection Application ...	Fitrianah et al.	2020
Music interest classification of Twitter users us ...	Yusra et al.	2017
Lexical based sentiment analysis - Verb, adverb ...	Shamsudin et al.	2016
Construction of the Malay language psychometric ...	Ahmad et al.	2017
Hate speech detection in Indonesian language on ...	Bunga Batara et al.	2019
Hate speech detection in Indonesian language on ...	Erizal et al.	2019
Hate speech detection on Indonesian Instagram ...	Pratiwi et al.	2019
Hate speech detection in Indonesian language In ...	Putra and Nurjanah	2020
Hate speech detection in Indonesian language on ...	Briliani et al.	2019
Recognizing the sarcastic statement on WhatsApp ...	Afiyati et al.	2018
Construction of Malay abbreviation corpus based ...	Omar et al.	2017
Context-sensitive normalization of social media ...	Kusumawardani et al.	2018
A taxonomy of Malay social media text	Maskat and Munarko	2019
Detecting opinion spams through supervised ...	Hazim et al.	2018
The development of Bahasa Indonesia corpora for ...	Jambak and Setiawan	2018
Review on sentiment analysis approaches for so ...	Abdullah et al.	2017
Sentiment Analysis of Noisy Malay Text: State ...	Abu Bakar et al.	2020
Emotion detection of tweets in Indonesian lan ...	Cahyaningtyas et al.	2017
Bias aware lexicon-based Sentiment Analysis of ...	Hijazi et al.	2017
Translated vs non-translated method for multilin ...	Ibrohim and Budi	2019b
Classification and quantification of user's emo ...	Jamaluddin et al.	2017
A hybrid model for social media sentiment analy ...	Putra et al.	2018b
Natural language processing based features for ...	Suhaimin et al.	2017
Modified framework for sarcasm detection and ...	Suhaimin et al.	2019
Concerns of thalassemia patients, carriers, and ...	Phang et al.	2021

End of Table