IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words

Haryo Akbarianto Wibowo* **Made Nindyatama Nityasya*** **Afra Feyza Akyürek***
Suci Fitriany* **Alham Fikri Aji*** **Radityo Eko Prasojo*•** **Derry Tanti Wijaya***
* Kata.ai Research Team, Jakarta, Indonesia
* Department of Computer Science, Boston University
• Faculty of Computer Science, Universitas Indonesia
{haryo,made,suci,aji,ridho}@kata.ai

Abstract

Indonesian language is heavily riddled with colloquialism whether in written or spoken forms. In this paper, we identify a class of Indonesian colloquial words that have undergone morphological transformations from their standard forms, categorize their word formations, and propose a benchmark dataset of Indonesian Colloquial Lexicons (IndoCollex) consisting of informal words on Twitter expertly annotated with their standard forms and their word formation types/tags. We evaluate several models for character-level transduction to perform morphological word normalization on this testbed to understand their failure cases and provide baselines for future work. As IndoCollex catalogues word formation phenomena that are also present in the non-standard text of other languages, it can also provide an attractive testbed for methods tailored for cross-lingual word normalization and non-standard word formation.

1 Introduction

Indonesian language is one of the most widely spoken languages in the world with around 200 million speakers. Despite its large number of speakers, in terms of NLP resources, Indonesian language is not very well represented (Joshi et al., 2020). Most of its data are in the form of unlabeled web and user generated contents in online platforms such as social media, which are noisy and riddled with colloquialism which poses difficulties for NLP systems (Baldwin et al., 2013a; Eisenstein, 2013a).

Traditionally, the majority of Indonesian colloquial or informal lexicons are borrowed words from foreign or local dialect words, and sometimes with phonetic and lexical modifications.[1] Increasingly however, Indonesian colloquial words are more commonly a morphological transformation[2] of their standard counterparts.[3] Despite these evolving lexicons, existing research on Indonesian word normalization has largely (1) relied on creating static informal dictionaries (Le et al., 2016), rendering normalization of unseen words impossible, and (2) for the specific task of sentiment analysis (Le et al., 2016) or machine translation (Guntara et al., 2020), with no direct implication to word normalization in general. Given the obvious utility of creating NLP systems that can normalize Indonesian informal data, we believe that the bottleneck is that there is no standard open testbed for researchers and developers of such system to test the effectiveness of their models to these colloquial words.

In this paper, we introduce IndoCollex, a new, realistic dataset aimed at testing normalization models to these phenomena. IndoCollex is a professionally annotated dataset, where each informal word is paired with its standard form and expertly annotated with its word formation type. The words are sampled from Twitter across different regions, therefore contain naturally occurring Indonesian colloquial words.

We benchmark character-level sequence-to-sequence transduction with LSTM (Deutsch et al., 2018; Cotterell et al., 2018) and Transformer (Vaswani et al., 2017) architectures, as well as a rule-based approach (Eskander et al., 2013; Moeljadi et al., 2019) on our data to understand their success and failure cases (§7.2, §7.3) and to provide baselines for future work. We also test methods for data augmentation in machine translation (back-translation), which to the best of our knowledge has never been applied to

[1] For example, *gue*, a common informal form of *aku* ('I', 'me'), is a word that originates from the Betawi dialect.

[2] We used the term morphological transformations broadly here to include word form changes at the respective interfaces of grammar (phonology, syntax, and semantics), following the definition by Trips (2017).

[3] For example, *laper*, a common informal form of *lapar* ('hungry'), is a phonetic change from its standard form.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3170–3183

character-level morphological transformation, and observe that adding back-translated data to train transformer improves its performance for normalizing informal words. We also test models in the other direction: generating informal from formal words, which can be useful for generating possible lexical replacements to standard text (Belinkov and Bisk, 2018).

2 Related Work

With the advent of social media and other user generated contents on the web, non-standard text such as informal language, colloquialism and slang become more prevalent. Concurrently, the rise of technologies like unsupervised language modeling opened up a new avenue for low-resource languages which lack annotated data for supervision. These systems typically only require large amounts of unlabeled text to train (Lample and Conneau, 2019; Brown et al., 2020). However, even when NLP systems require only unlabeled data to train, the varying degrees of formalism between different sources of monolingual data pose domain adaption challenges to NLP systems which are trained on one source (e.g. Wikipedia) to transfer to another (e.g. social media) (Eisenstein, 2013b; Baldwin et al., 2013b; Belinkov and Bisk, 2018; Pei et al., 2019). Worse yet, for an overwhelming majority of lower resource languages, unstructured and unlabeled text on the Internet is often the sole source of data to train NLP systems (Joshi et al., 2020). Therefore, addressing the formalism discrepancy will augment the types of web texts which can be employed in language technologies, especially for languages such as Indonesian which are subject to a high degree of informalism as will be discussed.

While this motivates research on training systems that are robust to non-standard data (Michel and Neubig, 2018; Belinkov and Bisk, 2018; Tan et al., 2020b,a), one intuitive direction is to normalize colloquial language use. Most of the work on colloquial language normalization has been done at the sentence-level: for colloquial English (Han et al., 2013; Lourentzou et al., 2019), Spanish (Cerón-Guzmán and León-Guzmán, 2016), Italian (Weber and Zhekova, 2016), Vietnamese (Nguyen et al., 2015), and Indonesian (Barik et al., 2019; Wibowo et al., 2020). However, research on the linguistic phenomena of non-standard text (Mattiello, 2005), which argues that slang words exhibit extra-grammatical morphological properties (such as portmanteaus, clipping) that distinguish them from the standard form, justifies the need for word-level normalization.

Word-level normalization also has its merit because due to its much lower hypothesis space, models can be trained using significantly smaller amount of data (e.g., compare SIGMORPHON's 10k examples to WMT's 10^6 at high-resource setting). Further, from our manual analysis of the top-10k most frequent Indonesian informal words we collected from Twitter, we find that around 95% of these words do not require context to normalize. Additionally, previous works such as Kulkarni and Wang (2018) have suggested that creating computational models for this generation of informal words can give us insights into the generative process of word formation in non-standard language. This is important because studies into the generative processes of word formation in non-standard text can deepen our understanding of non-standard text. Moreover, they are potentially applicable to many languages since word formation patterns are shared across languages (Štekauer et al., 2012), e.g., portmanteaus (such as *brexit*) have been found not only in English but also in many other languages such as Indonesian (Dardjowidjojo, 1979), Modern Hebrew (Bat-El, 1996), and Spanish (Piñeros, 2004). Finally, the studies may have broader applications including development of rich conversational agents and tools like brand name generators and headlines (Özbal and Strapparava, 2012).

Previous work that qualitatively catalogues or creates computational models for informal word formations such as shortening has mostly been in English, using LSTMs (Gangal et al., 2017; Kulkarni and Wang, 2018) or finite state machines (Deri and Knight, 2015) to generate informal words given the standard forms and the type of word formation. Most of the dataset: formal-informal word pairs labeled with their word formation used to train these models are also in English. Other dictionaries of informal English words include SlangNet (Dhuliawala et al., 2016), SlangSD (Wu et al., 2018), and SLANGZY (Pei et al., 2019). There is also a dataset that contains pairs of formal-informal Indonesian words (Salsabila et al., 2018), but they are not annotated with word formation mechanisms. To the best of our knowledge, ours is the first dataset of formal-informal lexicon in a language other than English that is annotated with their word formation types.

3 Indonesian Colloquialism

3.1 Indonesian Colloquial Words

Language evolves over time due to the process of language learning across generations, contact with other languages, differences in social groups, and rapid casual usages (Liberman et al., 2003). Each of these factors exists to a high degree in Indonesia, resulting in the constant evolution of its language due to contacts with over 700 local languages (Simons and Fennig, 2017), socioeconomic and education inequalities that result in varying level of adoption of the standard Indonesian (Azzizah, 2015), and the rise of social media usages with widespread *celeb* culture (Suhardianto et al., 2019; Heryanto, 2008) that causes new words to be invented and spread rapidly.

We catalog the following word formation types that are common in colloquial Indonesian.

1. **Disemvoweling**: elimination of some or all the vowels, e.g: *jangan* to *jgn* ('no' or 'don't'). Disemvoweling does not correspond to any phonetic change,
2. **Shortening or Clipping**: syllabic shortening of the original word, e.g: *internet* to *inet*. Unlike disemvoweling, shortening does imply phonetic change,
3. **Space/dash removal**: shortened version of writing Indonesian plural form, e.g.: *teman-teman* to *temanteman* or *teman2* ('friends'),
4. **Phonetic (sound) alteration**: slight change both in sound and spelling in text, but the number of syllables stay the same, e.g: *pakai* to *pake* or *pakek* ('use'),
5. **Informal affixation**: modification, addition or removal of affixes, e.g: *mengajari* to *ngajarin* ('to teach'),
6. **Compounding and acronym**: syllabic and letter compounds of one or more words akin to acronyms, abbreviations, and portmanteau, e.g: *anak baru gede* to *abg* ('teen'), *budak cinta* to *bucin* (literally, 'being a slave to love'),
7. **Reverse**: letter reversal, or colloquially known as "Boso Walikan" (Hoogervorst, 2014), e.g: *malang* (the name of a city in Indonesia) to *ngalam*.
8. **Loan words**: borrowed words, often from local language or English, e.g: *bokap* ('dad' in Betawi)
9. **Jargon**: tagline, terms that have been made into a popular term, e.g: *meneketehe*, from *mana aku tahu* (a jargon for 'how should I know?').

Some of the above transformations are also found in the literature of other languages, such as English and Korean. In English, disemvoweling was common during the texting (SMS) era in order to write faster and to save on message lengths e.g., *c u l8r* ('see you later'). Informal affixation (*cryin*, *sweet-ass*), compounding and portmanteaus (*btw*, *sexting*), and phonetic alteration (*dis is da wae*) are also present. In Korean, some compounded or shortened version of *Konglish* is also widely used (Khan and Choi, 2016), e.g., *chimaek* from *chicken* and *maek* ('beer'). Any insight we obtain through evaluating models on our dataset may therefore be of interest to other languages that share similar colloquial transformations; insights that may be increasingly paramount due to the rising prevalance of non-standard text in many languages on the web (Kulkarni and Wang, 2018; Joshi et al., 2020) and the challenges they pose to NLP systems (Belinkov and Bisk, 2018; Pei et al., 2019).

Loan word transformations that come from other languages require multilingual dictionaries/embeddings to normalize while jargons often require background knowledge. Aside from these two, we follow the previous work and hypothesize that the word formations that fall in other categories are mostly morphological transformations that can be learned at character-level (Kulkarni and Wang, 2018; Gangal et al., 2017). In §4, we describe how we curate this colloquial transformation data.

3.2 Indonesian Colloquialism Analysis

In this section, we motivate the importance of research on Indonesian colloquialism by highlighting their prevalence in Indonesian web text. We indeed observe that in its daily use Indonesians use colloquial Indonesian to generate contents in the web with (1) vocabularies that are different from formal Indonesian and (2) at a higher rate than colloquial use in the English language.

To compare colloquial and formal Indonesian (from Twitter and Lazada product reviews[4] and from Kompas news articles respectively (Tala, 2003)), we compute these dataset perplexities as well as their out-of-vocabulary (OOV) rates with respect to an Indonesian formal lexicon constructed from tokenizing Indonesian Wikipedia articles. For a fair comparison, we sample 3685 sentences from

[4] www.kaggle.com/grikomsn/lazada-indonesian-reviews

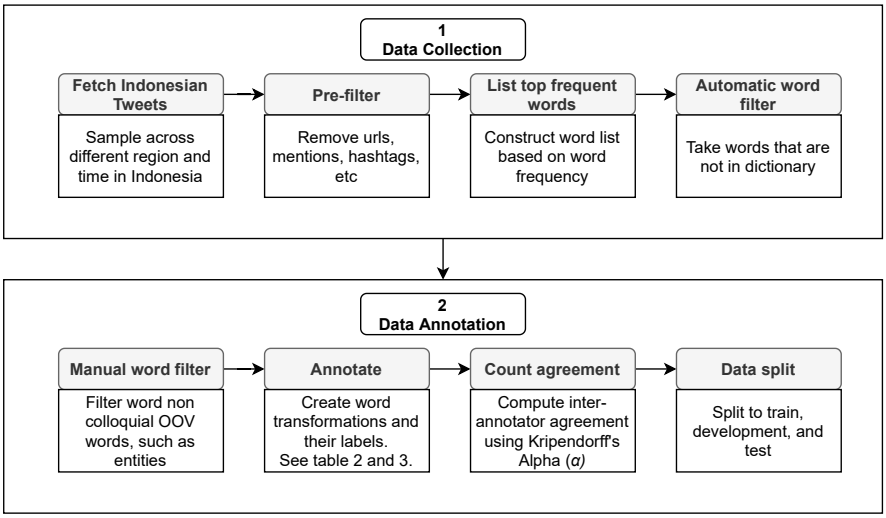

Figure 1: The data construction process composed of Data Collection and Data Annotation

each dataset based on the size of the smallest dataset. To compare to colloquial use in the English language, we also compare English tweets to an English formal lexicon constructed from English Wikipedia articles. We use Wikipedia to construct these lexicons to include named entities which are not typically present in traditional dictionaries.

Table 1 shows the OOV rate of the various datasets. Our OOV count excludes Twitter usernames, hashtags, mentions, URLs, dates, and numbers. To avoid rare words being captured as OOV, we also remove any token that only occurred once (shown as OOV-2) on the Table. We observe that the OOV rate of colloquial Indonesian is double the OOV rate of informal English. OOV of the formal Indonesian text (Kompas news) is low, as expected.

We use perplexity as a measure of impact of colloquialism beyond vocabulary usage and utilize a pre-trained Indonesian GPT-2 trained on Wikipedia[5] and Open AI's GPT-2[6] to calculate Indonesian and English data perplexities, respectively. Table 1 shows these perplexities.

Indonesian tweets have comparable perplexity as Lazada as they both use colloquial language. Both also have much higher perplexities than Kompas, implying that Indonesian LM finds that colloquial Indonesian is different than formal Indonesian. Similarly, English tweets have a higher perplexity compared to English Wikipedia (Radford et al., 2019). Notably, aside from Indonesian Twitter having around two times higher OOV rates: as high as 14.6% in OOV and 8.3% in OOV-2 than English Twitter, its perplexity too is significantly higher than English Twitter —suggesting that the non-standard word formation is a much more prominent issue when it comes to Indonesian, yet remains significantly under-researched.

Dataset	Lang.	OOV	OOV-2	Ppl
Twitter	ID	14.6%	8.3%	1617.0
Lazada review	ID	9.1%	7.0%	1824.3
Kompas news	ID	1.5%	1.1%	145.8
Wikipedia	ID	n/a	n/a	29.9
Twitter	EN	6.4%	4.5%	611.2
Wikipedia	EN	n/a	n/a	29.4

Table 1: OOV rates of Indonesian and English datasets

4 Data Collection and Annotation

Our dataset is constructed and manually annotated from a list of informal words obtained from Twitter. The data construction process is summarized in Figure 1. As an archipelago country, Indonesia is very diverse in local languages, which affects the way people use the Indonesian language. Hence, we sample 80 tweets per-day from March 2017 to May 2020, from each of the 34 provinces in Indonesia. We then select top 10k frequent tokens not appearing in our Wikipedia-based formal word dictionary and treat them as informal. Then we manually filter out from this list, OOV words that are not informal words such as product names or entities. Despite being sampled according to geolocation, we note that most of the informal words are more inclined to informal words commonly used in Jakarta. We suspect this is because Jakarta, being the center of Indonesian economy and pop culture (CITE), heavily influences the other regions through mainstream media. Further investigation on this aspect

[5]huggingface.co/cahya/gpt2-small-indonesian-522M
[6]huggingface.co/gpt2

is necessary and we leave this as a future work.

We assign four Indonesian native speakers[7], with formal education in linguistics and/or computational linguistics, to annotate each informal word with its standard form and label the pair with their word formation types according to our annotation codebook.[8] We annotate 9 different types of word formation mechanisms: disemvoweling, shortening, space/dash removal, phonetic (sound) alteration, affixation, compounding, reverse, loan word, and jargon. Since an informal word is often produced by stacking multiple transformations, we also annotate the transformation order, from the formal word to the informal. Some annotation examples are shown in Table 2. To simplify the transformation task, we assume single transformations and treat stacked transformations as a sequence of separate transformations. Words undergoing multiple transformations are broken down into different entries in our dataset. Ultimately, our dataset consists of parallel formal and informal Indonesian word pairs, each with its annotated word formation type from formal to informal. A sample of our dataset is shown in Table 3. Note that the same formal word with the same transformation may produce different informal words due to the open vocabulary of colloquial words.

Our dataset contains 3048 annotated word pairs[9] of which 2036 are those with morphological transformations (i.e., not loan words or jargons), which is comparable in size to other morphological transformation dataset such as the SIGMORPHON shared task (Cotterell et al., 2018). In comparison, Bengali, which is also a lower resource language comparable to Indonesian (Joshi et al., 2020), has 136 lemmas (and 4000 word forms) crowdsourced in the SIGMORPHON inflection dataset while our dataset has expertly annotated 1602 formal words (and 2036 informal variants).

In order to ensure the quality of our annotations, we sample 100 word pairs and compute Kripendorff's Alpha (α) (Hayes and Krippendorff, 2007) and Cohen's Kappa (κ) (Cohen, 1960) to measure agreement on word formation type annotations. The scores are $\alpha = 0.709$ $\kappa = 0.708$, showing that the annotators have substantial agreement on our dataset (Viera et al., 2005). We split the dataset into training, validation, and testing as in Table 4. Note that since reverse formation is quite rare, we augment the data and add additional reverse formation in the testing and validation sets.

In our experiments, we exclude loan word and jargon from the evaluation of character-level models, since these transformations are challenging, if not impossible to handle at the character-level alone without (1) additional resources such as multilingual dictionaries/embeddings and without (2) involving additional tasks such as translation.

5 Rule-Based Transformation Baseline

We believe that some formal to informal word formation mechanisms follow regular patterns. We manually define a rule-based system as one of our baselines (see Appendix). As we will demonstrate in the results section, there are several challenges entailed with a rule-based approach. Firstly, our rule-based transformation only works from formal to informal—as most of the colloquialism involves removing parts of the word, reverting from informal to formal Indonesian proves difficult for the rule-based system as it requires predicting the removed characters.

Secondly, the rule-based approach can not be universally applied. For example, in affixation, some Indonesian root words have sub-words similar to common morphological affixes in Indonesian such as *me-* or *-kan*. However, since these sub-words are part of the root words, they should not be removed/altered e.g., *membal* ('bouncy') cannot be transformed via informal affixation to *ngebal*, since *me-* in *membal* is part of the root word. Similarly, sound-alter transformation is applicable only to some words but not others e.g., *malam* ('night') can be altered to *malem*, but *galak* ('fierce') cannot be altered to *galek*. The rule of which words can be sound-altered seems arbitrary. In compounding, there is also no clear rule as to which abbreviation to use in different settings (e.g., *anak baru gede* is abbreviated to *ABG*, but *rapat kerja nasional* is abbreviated to *rakernas* instead of *RKN*). Lastly, as a single word may have multiple possible transformations that can apply, since rule-based system cannot rank these possible outputs, it randomly picks one of the candidates.

6 Character-Level Seq2Seq Models

Previous approaches for generating transformed words model the task as a character-level sequence-to-sequence (SEQ2SEQ) problem: the characters

[7] formally employed by our company, Kata.ai.
[8] https://github.com/haryoa/indo-collex
[9] Full dataset: https://github.com/haryoa/indo-collex

Informal Word	Annotated Formal	Annotated Word Formation Type
kuy	ayo (let's go)	original → sound-alter → reverse **note:** ayo → y*uk* → *kuy*
gpp	tidak apa-apa (no problem)	original → shortening → disemvoweling **note:** tidak apa-apa → *gapapa* → *gpp*
ngeselin	mengesalkan (annoying)	original → affixation → sound-alter **note:** mengesalkan → ngesal*in* → nges*e*lin

Table 2: Examples of informal words annotated with their formal versions, alongside the transformation sequences.

Source	Target	Word Formation Tag
ayo (formal of "let's go")	yuk (informal of "let's go")	sound-alter
ayo (formal of "let's go")	yuks (informal of "let's go")	sound-alter
yuk (informal of "let's go")	kuy (informal of "let's go")	reverse
yuks (informal of "let's go")	skuy (informal of "let's go")	reverse
kemarin (formal of "yesterday")	kmrn (informal of "yesterday")	disemvoweling
nasi goreng (fried rice)	nasgor	compounding
membuka (formal of "opening")	ngebuka (informal of "opening")	affixation

Table 3: Example entries in our colloquial transformation dataset.

Word Formation Tag	Train	Valid	Test
sound-alter	489	21	35
shorten	363	41	43
disemvoweling	323	30	31
affixation	165	30	20
space/dash removal	157	26	21
acronym	155	23	16
reverse	5	15	27
Total	1657	186	193

Table 4: Formal (F)→Informal (I) data distribution.

from the root word and an encoding of the desired transformation type are given as input to a neural encoder, and the decoder is trained to produce the transformed word, one character at a time (Gangal et al., 2017; Deutsch et al., 2018; Cotterell et al., 2017). In reality however, transformation types are often implied, but not given. For example, an Indonesian speaker will be able to transform the formal *tolong* ('help') to *tlg* given examples that *jangan* ('don't') can be transformed to *jgn*, even without the transformation type i.e., disemvoweling being specified. Thus, we also experiment with these SEQ2SEQ models for generating informal words from formal (and vice versa) without inputting any word formation tag to see if the models can induce the desired transformation type based on morphologically similar words in the training examples. We also use these models trained to generate outputs without word formation input to generate back-translated data to augment our training (§7.1).

6.1 BiLSTM

The dominant model for character-level transduction that have been applied to many tasks such as morphological inflection (Cotterell et al., 2017), morphological derivation (Deutsch et al., 2018), and informal word formation (Gangal et al., 2017) adopts a character-level SEQ2SEQ model that learns to generate a target word from its original form given the desired transformation. These models typically use bi-directional LSTM with attention (Luong et al., 2015) to learn these transformations as orthographic functions. For the task of morphological derivation, the SOTA model (Deutsch et al., 2018) also proposes a dictionary constraint approach where the decoding process is restricted to output tokens listed in the dictionary, which improves the accuracy of their model.

We evaluate this SOTA character SEQ2SEQ that leverages dictionary constraint (BiLSTM+Dict), whose code is publicly available,[10] on our data. Following their approach, we train this model for 30 epochs with a batch-size of 5 using Adam optimizer with initial learning-rate of 0.005, an embedding size of 20, and a hidden state size of 40. For the dictionary constraint, we construct dictionaries of formal words from Indonesian Wikipedia (§3.2) and informal words we collected from Twitter (i.e., words we collected from Twitter that do not appear in our Wikipedia-based formal word dictionary §4).

6.2 Transformer

Given that more recently Transformer has been shown to outperform standard recurrent models on several character-level transduction tasks including morphological inflection and historical text normalization, grapheme-to-phoneme conversion, and transliteration (Wu et al., 2020); we evaluate character-based Transformer model (Vaswani et al., 2017) on our dataset. We conduct hyperparame-

[10] github.com/danieldeutsch/derivational-morphology

ter tuning on the size of the character embeddings, the number of layers, and the number of attention heads of the Transformer. For training, we use Adam with an initial learning rate of 0.005, a batch size of 128 (following (Wu et al., 2020)), and train for a maximum of 200 epochs, returning the model with the least validation loss.

7 Experiment and Results

We evaluate standard character-level transduction models on our dataset to assess its difficulty. Our goal is not to train SOTA models for word normalization but rather to test these models for such task on our data, and elucidate what features of the data make it difficult.

7.1 Experiment Settings

We train and evaluate the BiLSTM+Dict and Transformer models on our dataset. The models are trained and evaluated in both directions: formal↔informal (F↔I) Indonesian. However, as mentioned previously, we only explore formal→informal (F→I) for the rule-based model. We also train the SEQ2SEQ models with and without inputting the word formation tag. Each experiment took about 3 hours on a K80 GPU.

Aside from training the models to transform formal↔informal words, we also use the Transformer model to predict the word formation tag $t \in T$, where T is the set of word formation types in our dataset, that best applies given an informal word and its corresponding formal form (I→F) or vice versa (F→I) (i.e., Transformer$_{(I \rightarrow F) \rightarrow T}$ and Transformer$_{(F \rightarrow I) \rightarrow T}$).

We experiment with using backtranslation (Sennrich et al., 2016), which has been used to learn novel inflections in statistical machine translation (Bojar and Tamchyna, 2011), at the character-level to increase the training data for I→F. Using Transformer$_{F \rightarrow I}$ model that performs best on the validation set, we generate informal words from the words in our formal dictionary sorted by frequencies. We experiment with generating $M = kN$ additional word pairs, where $k = \{1, 2, 3\}$ and N is the number of word pairs in the original training data. We similarly augment training data for F→I by using the Transformer$_{I \rightarrow F}$ model that performs the best on the validation set to generate formal words from our informal word dictionary.

To ensure that the augmented data has similar transformation distribution as the original training data, we predict the word formation type that best applies to each generated word pair using the Transformer$_{(I \leftrightarrow F) \rightarrow T}$ model that performs best on validation. For each word formation type, we add rM generated pairs with such type to our training data based on its ratio r in the original training.

Each model's performance is measured by the top-*1* and top-*10* accuracy. Since formal→informal transformation is rather flexible, we also capture the BLEU score of the model's output. We report performances of the hyperparameter-tuned models that perform best on the validation set.

7.2 Results

Our experiment results are shown in Table 5. Generally, Transformer models outperform all other models. Specifying the target word formation type improves the performance of both models. Backtranslation is also shown to improve the performance of the Transformer. Transformer with added backtranslation and word formation tag yields the best test performance in both directions.

We also observe that in average the performance of the models are higher in the I→F direction than F→I. We observe similar trends when predicting word formation types given word pairs. The accuracy of the Transformer$_{(I \rightarrow F) \rightarrow T}$ model that predicts the type that applies given an informal word and its corresponding formal form is 81.4%; which is significantly higher than the 65.0% accuracy of the Transformer$_{(F \rightarrow I) \rightarrow T}$ model that predicts the type given a formal word and its corresponding informal form. This may point to the inherent ambiguity of generating informal words from the formal words. Due to the open-vocabulary of informal words, there are potentially many ways to transform a formal word into informal forms.

Surprisingly, rule-based transformation outperforms BiLSTM+Dict and several non-optimal Transformer configurations in terms of top-*1* accuracy. However, rule-based transformation does not perform well in terms of top-*10* accuracy. We observe that the rule-based transformation does not always manage to produce 10 transformation candidates, therefore missing out on the extra chances to correctly guess the output.

7.3 Discussion

In this section, we discuss failures and success cases of the best performing model (Transformer) on our dataset, elucidate what the model learns,

Model	Informal to Formal					Formal to Informal				
	Dev		Test			Dev		Test		
	Top1	BLEU	Top1	Top10	BLEU	Top1	BLEU	Top1	Top10	BLEU
Rule-Based	-	-	-	-	-	27.9	43.9	34.7	53.4	50.2
BiLSTM + Dict	23.5	53.0	30.5	58.8	57.9	18.8	47.3	22.8	56.1	47.7
BiLSTM + Dict + word-formation tag	25.7	54.8	30.5	56.3	53.2	30.6	60.8	30.1	62.5	54.4
Transformer	30.1	59.3	27.9	60.6	61.6	19.4	53.3	21.2	56.5	48.7
Transformer + word-formation tag	**33.9**	64.4	35.8	65.9	61.6	31.2	59.2	22.3	54.2	48.2
Transformer + BT	31.7	65.7	32.1	66.4	**63.7**	22.6	57.1	24.4	58.4	53.4
Transformer + BT + word-formation tag	33.3	**66.5**	**37.4**	**70.2**	62.1	**36.6**	**69.2**	**35.8**	**67.5**	**57.0**

Table 5: Experiment Result for Informal and Formal Colloquial transformation.

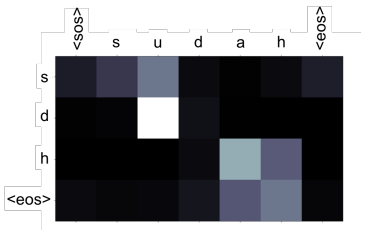

Figure 2: Attention matrix of *sudah* (F) → *sdh* (I) *without* word formation tag (column: source, row: target). The model learns to disemvowel implicitly by paying attention to the vowels and removing them.

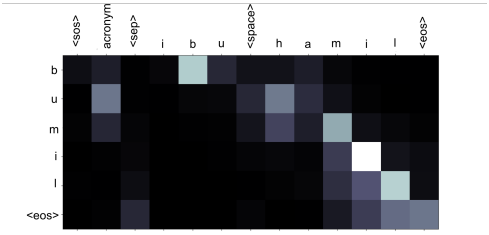

Figure 3: Attention matrix of *ibu hamil* F→I transformation *with* word formation tag (column: source, row: target). The model pays attention to the tag (acronym) while getting the prefix *bu-* from the first word and the suffix *-mil* from the second.

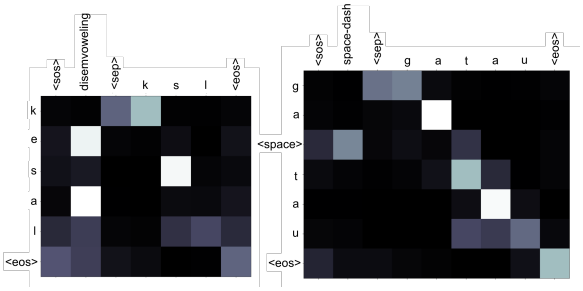

Figure 4: Attention matrix of *ksl* (I) → *kesal* (F) and *gatau* (I) → *ga tau* (F) *with* tag (column: source, row: target). The model learns to pay attention to the tag while regenerating the missing vowels and space.

and analyze features of the data that make it challenging for the model. As seen in Table 5, when the desired word formation is not given, the Transformer has worse performance when performing F→I transformation compared to I→F. This is because transforming from formal to informal has a higher level of ambiguity i.e., a word can be made informal by multiple possible word formations.

If the word formation type is not given, we observe that Transformer will learn to select the type implicitly. For example, it selects the disemvoweling mechanism implicitly as it pays attention to vowels in the word while removing them e.g., to correctly generate the informal *sdh* from the formal *sudah* (meaning, 'already') Figure 2). If the input consists of two words (separated by space), the model assumes the space/dash removal mechanism, paying attention to the characters before and after the space while removing the space e.g., given the word *ga tau* (meaning, 'don't know'), the model removes the space and correctly returns *gatau*.

However, the Transformer may select an incorrect transformation when the target word formation is not given e.g., the phrase *ibu hamil* ('pregnant mother') is often expressed as *bumil* (acronym). Without tag, the model performs a space/dash removal instead, and produced incorrect *ibuhamil*. Figure 3 shows how the model attends to the tag when it is given and applies the correct mechanism.

We observe that the model also attends to the tag when transforming the word in the reverse (I→F) direction e.g., the model pays attention to the tag while correctly generating the vowels of a disemvoweled words *ksl* to *kesal* ('annoyed') or the space between the compounded word *gatau* to *ga tau* (Figure 4).

In general, we observe that formal to informal transformation is challenging, since multiple valid informal words are possible even for a given word and word formation type. For example, *kamu* ('you') can be written informally as *km* or *kmu* both with the same disemvoweling transformation. Some word formation mechanisms are also ambiguous. For example, *budak cinta*'s acronym is *bucin* (using the *prefix* of the second word), whereas *ibu hamil*'s acronym is *bumil* (using the *suffix* of the second word). The acronym transformation seems to

be applied on a case-by-case basis with no clear pattern. Reversing acronym to its original phrases is even more challenging (with or without tags) since it requires models to reconstruct the full phrase given minimum context e.g., reconstructing *anak layangan* ('tacky') from its acronym *alay*.

Another challenging transformation is affixation. Since *me-* and its different variants (*mem-*, *men-*, etc.) are common morphological prefixes in Indonesian, we observe that our best model, the Transformer, often puts *me-* in I→F affixation transformation, mistakenly transforming for example, *nyantai* ('to relax') into *menyantai* (expected: *bersantai*). This suggests that more training data may be needed to capture various affixation.

On the other hand, in sound alteration, we observe that Transformer successfully learns to sound-alter even when the word formation is not explicitly mentioned. For example, it learns to transform the informal *pake* ('to wear') to *pakai* (attending to the characters *e* when outputting *ai*), *kalo* ('if') to *kalau* (attending to the character *o* when outputting *au*), and *mauuu* ('want') to *mau* (attending to the characters *uuu* when outputting *u*).

8 Ethical Consideration

Normalizing informal Indonesian language might serve as a bridge to connect the generational gap in the use of the language, as the informal Indonesian language is more popular among the younger populace. Furthermore, it can potentially bridge linguistic differences across the Indonesian archipelago. Although we attempt to collect informal data from each province in Indonesia, the resulting informal dataset is still mostly Jakarta-centric, and further scraping and verification of the linguistic coverage is necessary for future work. Finally, as not every Indonesian speaks perfect standard Indonesian, having an NLP interface (such as chatbots) that can readily accept (process and understand via normalization) any kind of informality that might arise promotes inclusivity that all NLP research should strive for.

9 Conclusion and Future Work

We show that colloquial and formal Indonesian are vastly different in terms of OOV-rate and perplexity, which poses difficulty for NLP systems that are trained on formal corpora. This significant gap between train and test sets in terms of formalism may hinder progress in Indonesian NLP research. We propose a new benchmark dataset for Indonesian colloquial word normalization that contains formal-informal word pairs annotated with their word formation mechanisms. We test several dominant character-level transduction models as baselines on the dataset and observe that different word formation mechanisms pose different levels of difficulties to the models with transformation to informal forms being more challenging due to the higher degree of transformation variants. Through this dataset, we intend to provide a standard benchmark for Indonesian word normalization and foster further research on models, datasets and evaluation metrics tailored for this increasingly prevalent and important problem.

In the future, we are interested to use the context in which the words occur, either textual (e.g., sentences) or other modalities (e.g., images or memes), to improve word transformation (formal ↔ informal) by using the context as either implicit signal (Wijaya et al., 2017) or explicit signal for "translating" between the formal and informal word forms based on similarities between their sentence contexts (Feng et al., 2020; Reimers and Gurevych, 2020) or image contexts (Bergsma and Van Durme, 2011; Kiela et al., 2015; Hewitt et al., 2018; Khani et al., 2021). We are also interested to learn if simple clustering of contexts within which the words occur can help us learn the mapping between the formal and informal words similar to finding paraphrase matching (Wijaya and Gianfortoni, 2011). Lastly, we are interested in the use of text normalization to augment data for training informal text translation (Michel and Neubig, 2018; Jones and Wijaya, 2021) or for training other downstream applications such as framing identification (Card et al., 2015; Liu et al., 2019; Akyürek et al., 2020), which are typically trained on formal news text, on informal social media text.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful comments. Derry Tanti Wijaya is supported in part by the U.S. NSF grant 1838193 (BIGDATA: IA: Multiplatform, Multilingual, and Multimodal Tools for Analyzing Public Communication in over 100 Languages), DARPA HR001118S0044 (Learning with Less Labeling program), and the Department of the Air Force FA8750-19-2-3334 (Semi-supervised Learning of Multimodal Representations).

References

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Yuni Azzizah. 2015. Socio-economic factors on indonesia education disparity. *International Education Studies*, 8(12):218–229.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013a. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013b. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

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, Hong Kong, China. Association for Computational Linguistics.

Outi Bat-El. 1996. Selecting the best of the worst: the grammar of hebrew blends. *Phonology*, 13(3):283–328.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1764. Citeseer.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Jhon Adrián Cerón-Guzmán and Elizabeth León-Guzmán. 2016. Lexical normalization of spanish tweets. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 605–610.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. *arXiv preprint arXiv:1708.09151*.

Soenjono Dardjowidjojo. 1979. Acronymic patterns in indonesian. *Pacific Linguistics Series C*, 45:143–160.

Aliya Deri and Kevin Knight. 2015. How to make a frenemy: Multitape fsts for portmanteau generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–210.

Daniel Deutsch, John Hewitt, and Dan Roth. 2018. A distributional and orthographic aggregation model for english derivational morphology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1938–1947.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Slangnet: A wordnet like resource for english slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4329–4332.

Jacob Eisenstein. 2013a. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Jacob Eisenstein. 2013b. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing spontaneous orthography. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 585–595.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Varun Gangal, Harsh Jhamtani, Graham Neubig, Eduard Hovy, and Eric Nyberg. 2017. Charmanteau: Character embedding models for portmanteau creation. *arXiv preprint arXiv:1707.01176*.

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.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):1–27.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Ariel Heryanto. 2008. *Popular culture in Indonesia: Fluid identities in post-authoritarian politics*. Routledge.

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.

Tom G Hoogervorst. 2014. Youth culture and urban pride: The sociolinguistics of east javanese slang. *Wacana*, 15(1):104–131.

Alex Jones and Derry Tanti Wijaya. 2021. Sentiment-based candidate selection for nmt. *arXiv preprint arXiv:2104.04840*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Irfan Ajmal Khan and Jin-Tak Choi. 2016. Lexicon-corpus based korean unknown foreign word extraction and updating using syllable identification. *Procedia Engineering*, 154:192–198.

Nikzad Khani, Isidora Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. Cultural and geographical influences on image translatability of words across languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 198–209, Online. Association for Computational Linguistics.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal. Association for Computational Linguistics.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1424–1434.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Mark Liberman, Yoon-Kyoung Joh, John Laury, and Marjorie Pak. 2003. Types of language change. `http://web.archive.org/web/20200610043656/https://www.ling.upenn.edu/courses/Fall_2003/ling001/language_change.html`. [Online; archived 2020-06-10].

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 335–345.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Elisa Mattiello. 2005. The pervasiveness of slang in standard and non-standard english. *Mots Palabras Words*, 6:7–41.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

David Moeljadi, Aditya Kurniawan, and Debaditya Goswami. 2019. Building cendana: a treebank for informal indonesian.

Vu H Nguyen, Hien T Nguyen, and Vaclav Snasel. 2015. Normalization of vietnamese tweets on twitter. In *Intelligent Data Analysis and Applications*, pages 179–189. Springer.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 703–711, Jeju Island, Korea. Association for Computational Linguistics.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889.

Carlos-Eduardo Piñeros. 2004. The creation of portmanteaus in the extragrammatical morphology of spanish. *Probus*, 16(2):203–240.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Nikmatun Aliyah Salsabila, Yosef Ardhito Winatmoko, Ali Akbar Septiandri, and Ade Jamal. 2018. Colloquial indonesian lexicon. In *2018 International Conference on Asian Language Processing (IALP)*, pages 226–229. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Gary F Simons and Charles D Fennig. 2017. *Ethnologue: languages of Asia*. SIL International Dallas.

Pavol Štekauer, Salvador Valera, and Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: A typological survey*. Cambridge University Press.

Suhardianto Suhardianto et al. 2019. Colloquail, slang and transformational language: Comperative study. *JURNAL BASIS*, 6(1):105–118.

Fadillah Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020a. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020b. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Carola Trips. 2017. Morphological change. In *Oxford Research Encyclopedia of Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Daniel Weber and Desislava Zhekova. 2016. Tweetnorm: text normalization on italian twitter data. In *Proceedings of the 13th Conference on Natural Language Processing*, pages 306–312.

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. In *2020 International Conference on Asian Language Processing (IALP)*, pages 310–315. IEEE.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463.

Derry Tanti Wijaya and Philip Gianfortoni. 2011. " nut case: what does it mean?" understanding semantic relationship between nouns in noun compounds through paraphrasing and ranking the paraphrases. In *Proceedings of the 1st international workshop on Search and mining entity-relationship data*, pages 9–14.

Liang Wu, Fred Morstatter, and Huan Liu. 2018. Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52(3):839–852.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv e-prints*, pages arXiv–2005.

A Rule-based transformation

Our rule-based transformation can be described in Table 6.

Condition	Transformation	Example
Type: Affixation		
prefix meng- followed by a consonant	replace prefix to nge-	menghina - ngehina
prefix menc-	replace prefix to ny-	mencari - nyari
prefix mem- followed by a consonant	replace prefix to nge-	membuka - ngebuka
prefix men- followed by a consonant	replace prefix to nge-	menjitak ->ngejitak
prefix me- followed by l, q, r, w	replace prefix to nge-	melempar ->ngelempar
suffix -i	replace suffix to -in	pukuli ->pukulin
suffix -kan	replace suffix to -in	hidangkan ->hidangin
Type: Shorten		
prefix me- followed by ng	remove me-	mengegas ->ngegas
prefix me- followed by ny	remove me-	menyanyi ->nyanyi
prefix me- followed by m + vowel	remove me-	memukul ->mukul
prefix me- followed by n + vowel	remove me-	menendang ->nendang
prefix h-	remove h-	habis ->abis
identic duplicate words	replace word with 2	makan-makan ->makan2
Type: Sound-alteration		
last a	replace to e	malam - malem
last i	replace to e	kemarin ->kemaren
last ai	replace to e	sampai ->sampe
last au	replace to o	kalau-kalo
last ai	replace to ae	main - maen
last -nya	replace to -x	sepertinya - sepertix
last p	replace to b	mantap - mantab
last s	replace to z	habis - habiz
Compounding		
any pattern	select the first character	anak baru gede - abg
Second occurrence of cons. + vowel	All character before the pattern	butuh cinta - bucin
Second occurrence of cons. + vowel	All character up to the cons.	nasi goreng - nasgor
Disemvowelling		
any pattern	randomly remove vowels	kemarin - kmarin, kamu - km
Reverse		
any pattern	reverse the word	yuk - kuy

Table 6: List of rule-based transformation.