Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201961105
Vol. 6, No. 2, April 2019, hlm. 219-228 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

KOMBINASI METODE RULE-BASED DAN N-GRAM STEMMING UNTUK MENGENALI STEMMER BAHASA BALI

Made Agus Putra Subali[1], Chastine Fatichah[2]

[1,2]Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi,
Institut Teknologi Sepuluh Nopember, Surabaya
Email: [1]madeagusputrasubali@gmail.com, [2]chastine@if.its.ac.id

(Naskah masuk: 06 Oktober 2018, diterima untuk diterbitkan: 08 Januari 2019)

Abstrak

Proses untuk mengekstraksi kata dasar dari kata berafiks dikenal dengan istilah *stemming* yang bertujuan meningkatkan *recall* dengan mereduksi variasi kata berafiks ke dalam bentuk kata dasarnya. Penelitian terdahulu tentang *stemming* bahasa Bali pernah dilakukan menggunakan metode *rule-based*, tapi afiks yang diluluhkan hanya prefiks dan sufiks, sedangkan variasi afiks lain tidak diluluhkan, seperti infiks, konfiks, simulfiks, dan kombinasi afiks. Penelitian tentang *stemming* menggunakan pendekatan *rule-based* telah diterapkan di berbagai bahasa yang berbeda. Metode *rule-based* memiliki kelebihan jika diterapkan pada domain yang sederhana, maka *rule-based* mudah untuk diverifikasi dan divalidasi, tapi memiliki kelemahan saat diterapkan pada domain dengan level kompleksitas yang tinggi, apabila sistem tidak dapat mengenali *rules*, maka tidak ada hasil yang diperoleh. Untuk mengatasi kelemahan *stemming* menggunakan *rule-based*, kami menggunakan metode *n*-gram *stemming*, dimana kata berafiks dan kata dasar diubah ke bentuk *n*-gram, kemudian tingkat kemiripan antara *n*-gram kata berafiks dan *n*-gram kata dasar diukur menggunakan metode *dice coefficient*, apabila tingkat kemiripannya memenuhi nilai ambang batas yang ditentukan, maka kata dasar yang dibandingkan dengan kata berafiks ditampilkan. Pada penelitian ini, kami mengembangkan metode *stemmer* yang meluluhkan seluruh variasi afiks pada bahasa Bali dengan mengombinasikan pendekatan *rule-based* dan metode *n*-gram *stemming*. Berdasarkan pengujian yang telah dilakukan untuk kesepuluh *query* metode yang diusulkan memperoleh rerata akurasi *stemming* lebih baik 96,67% dari metode terdahulu 75%, sedangkan untuk kelima *query* metode *n*-gram *stemming* dapat mengenali beberapa kata berafiks diluar *rules*. Penelitian berikutnya, kami akan memperhatikan semantik setiap kata dan tahap validasi menggunakan aplikasi *text mining*.

Kata kunci: *stemmer bahasa Bali, rule-based stemming, n-gram stemming, dice coefficient.*

A COMBINATION OF METHODS RULE-BASED AND N-GRAM STEMMING TO RECOGNIZE BALINESE LANGUAGE STEMMER

Abstract

A process for extracting a stem word from the inflected word is known as stemming which aims to increase recall by reducing the variation of the inflected word into its stem word form. Previous research on stemming the Balinese language has been done using the rule-based method, but the affixes that are removed are only prefixes and suffixes, while other variations of affixes are not removed, such as infixes, confixes, simulfiks, and combinations of affixes. Research on stemming using the rule-based approach has been applied in a variety of different languages. The rule-based method has advantages when applied to a simple field, rule-based is easy to verify and validate, but has weaknesses when applied to domains with a high level of complexity, if the system cannot recognize rules, no results are obtained. To overcome the stemming weaknesses using rule-based, we use the n-gram stemming method, where the inflected word and stem word are converted to the n-gram form, then the level of similarity between the n-gram of the inflected word and the stem word is measured using the dice coefficient method, when the level of similarity meets the defined threshold value, then the stem word is displayed. In this study, we developed a stemmer method that removes all variations of affixes in the Balinese language by combining the rule-based approach and the n-gram stemming method. Based on the experiments for the ten queries the proposed method get 96,67% stemming accuracy than the previous method 75%, while for the five queries for the n-gram stemming method can recognize some inflected words outside the rules. The next study, we will pay attention to the semantics of each word and the validation stage using text mining application.

Keywords: *Balinese language stemmer, rule-based stemming, n-gram stemming, dice coefficient.*

1. PENDAHULUAN

Bahasa Bali merupakan bahasa ibu bagi masyarakat etnis Bali memiliki kedudukan dan fungsi yang sangat penting. Komunikasi keseharian etnis Bali sering didominasi oleh pemakaian bahasa Bali, terutama dalam topik pembicaraan yang bersifat tradisional, seperti membahas masalah adat, kebudayaan, dan agama. Dalam bahasa Bali dikenal adanya kata dasar dan kata turunan. Kata turunan sering disebut dengan istilah kata berimbuhan. Istilah imbuhan dapat disejajarkan dengan afiks. Kata berafiks dalam bahasa Bali dapat dibedakan menurut tempatnya melekat pada bentuk dasar atau asal, yaitu prefiks, sufiks, infiks, konfiks, simulfiks, dan kombinasi afiks (Granoka dkk, 1996).

Proses untuk mengekstraksi kata dasar dari kata berafiks dikenal dengan istilah *stemming* (Balasankar dkk, 2016) yang bertujuan meningkatkan *recall* dengan mereduksi variasi kata berafiks ke dalam bentuk kata dasarnya (Patil & Patil, 2017), (Pramudita dkk, 2018). Penelitian tentang *stemming* bahasa Bali pernah dilakukan oleh (Nata & Yudiastra, 2017), pada penelitian tersebut, peneliti menggunakan metode *rule-based* dengan mengadopsi algoritma (Tala, 2003), tapi afiks yang diluluhkan hanya prefiks dan sufiks, sedangkan variasi afiks lain tidak diluluhkan, seperti infiks, konfiks, simulfiks, dan kombinasi afiks.

Metode *rule-based* merupakan metode yang menggunakan *rules* sebagai representasi pengetahuan untuk diimplementasikan ke dalam sistem (Ligeza, 2006), (Nikolopoulos, 1997), (Lindsay, 1988). Metode *rule-based* sangat bergantung pada penalaran manusia sebagai *expert* dalam memecahkan masalah. Pendekatan *stemming* menggunakan *rule-based* telah diterapkan di berbagai bahasa yang berbeda (Memet dkk, 2017). Bahasa Inggris, (Lovins, 1968), (Porter, 1980), (Porter, 2001), (Paice, 1994), (Krovetz, 1993). Bahasa Arab, (De-Roeck & Al-Fares, 2000), (Larkey dkk, 2002). Bahasa Perancis, (Moulinier dkk, 2001). Bahasa Benggala, (Majumder dkk, 2007). Bahasa Turki, (Dincer & Karaoglan, 2003). Bahasa Indonesia, (Nazief & Adriani, 1996), (Adriani dkk, 2007).

Metode *rule-based* memiliki kelebihan jika diterapkan pada domain yang sederhana, maka *rule-based* mudah untuk diverifikasi dan divalidasi, tapi memiliki kelemahan pada saat diterapkan pada domain dengan level kompleksitas yang tinggi, apabila sistem *rule-based* tidak dapat mengenali *rules*, maka tidak ada hasil yang diperoleh (Grosan & Abraham, 2011) dan pendekatan *stemming* menggunakan *rule-based* hanya spesifik terhadap bahasa yang digunakan (Mayfield & McNamee, 2003). Untuk mengatasi kelemahan *stemming* menggunakan *rule-based*, kami menggunakan metode *n*-gram *stemming* yang dirancang oleh Adamson & Boreham, pendekatan ini menunjukan bahwa kata yang memiliki kesamaan struktural lebih tinggi cenderung sama dengan artinya (Adamson & Boreham, 1974), (Sembok & Bakar, 2011), dimana kata berafiks dan kata dasar diubah ke bentuk *n*-gram, kemudian tingkat kemiripan antara *n*-gram kata berafiks dan *n*-gram kata dasar diukur menggunakan metode *dice coefficient*, apabila memenuhi nilai ambang batas yang ditentukan, maka kata dasar yang dibandingkan dengan kata berafiks ditampilkan.

Pada penelitian ini, kami mengembangkan metode *stemmer* yang meluluhkan seluruh variasi afiks pada bahasa Bali dengan mengombinasikan pendekatan *rule-based* dan metode *n*-gram *stemming*. Untuk membuktikan metode yang diusulkan dapat memberikan hasil akurasi *stemming* yang lebih optimal, kami melakukan serangkaian pengujian, seperti membandingkan hasil *stemming* metode yang diusulkan dengan metode terdahulu, dimana dari sepuluh *query* yang diberikan metode yang diusulkan memperoleh akurasi *stemming* lebih baik 96,67% dibandingkan metode terdahulu 75%, menentukan nilai ambang batas yang paling sesuai pada seluruh variasi afiks, kisaran nilai ambang batas yang digunakan adalah 0,70, 0,65, dan 0,60, serta membandingkan beberapa variasi karakter *n*-gram pada metode *n*-gram *stemming*, yaitu *bi*-gram dan *tri*-gram.

2. METODE PENELITIAN

Penelitian ini terdiri dari beberapa tahapan antara lain studi literatur, penyusunan *list* kata dasar, perancangan sistem, serta hasil dan pembahasan.

2.1. Penyusunan *List* Kata Dasar

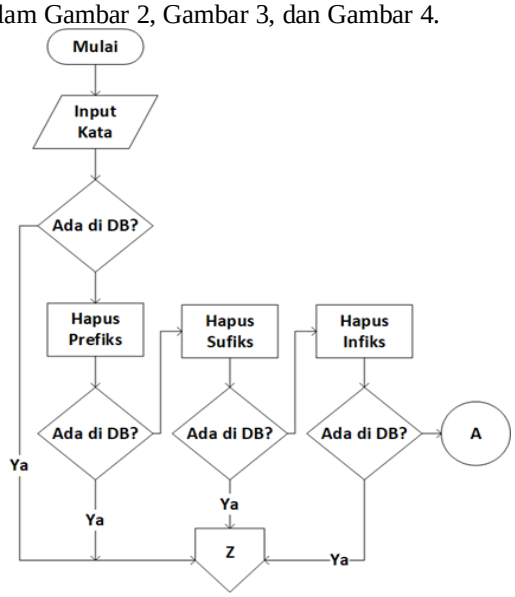
Sejumlah kata dasar bahasa Bali digunakan untuk memastikan bahwa hasil dari meluluhkan kata berafiks sesuai dengan bentuk dasarnya dan setiap kata dasar juga digunakan untuk dibandingkan dengan kata input pada metode *n*-gram *stemming*.

Penyusunan *list* kata dasar bahasa Bali diperoleh dari studi literatur pada buku dengan judul "*Tata Bahasa Baku Bahasa Bali*" dan artikel berita *online* bahasa Bali sejumlah lima puluh artikel dari situs *https://balitv.tv/category/news/orti-bali* selama bulan Nopember 2017, sejumlah kata berafiks yang diperoleh dari buku dan artikel berita bahasa Bali akan ditentukan bentuk kata dasarnya melalui penilaian yang diberikan oleh *expert* kemudian kata dasar disimpan dalam *database*. Kata dasar yang terkumpul sejumlah seribu kata dasar yang dapat diunduh melalui tautan *http://bitly.com/2DWditK*. Adapun tahapan mengumpulkan kata dasar dalam bahasa Bali dilakukan dalam Gambar 1.

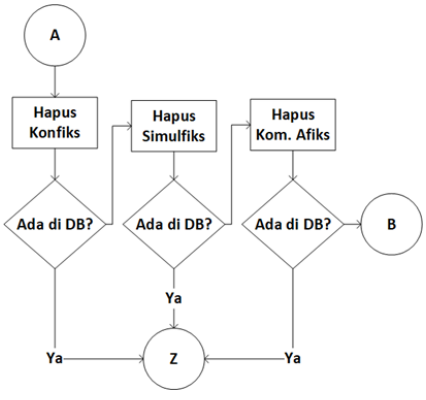
Gambar 1. Alur Proses Penyusunan *List* Kata Dasar

2.2. Perancangan Sistem

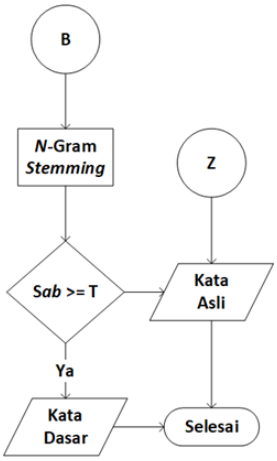
Pada penelitian ini, kami mengombinasikan metode *rule-based* dan *n*-gram *stemming* untuk *stemmer* bahasa Bali, dimana pada tahap awal dilakukan perbandingan apakah kata input merupakan kata dasar atau bukan dengan cara membandingkannya pada kamus kata dasar yang tersimpan pada *database*, apabila kata input merupakan kata dasar, maka kata dasar ditampilkan, apabila kata input bukan merupakan kata dasar, maka dilakukan proses meluluhkan afiks, dimulai dari meluluhkan prefiks, sufiks, infiks, konfiks, simulfiks, dan kombinasi afiks apabila *rules* yang tersedia tidak dapat mengenali kata input, maka untuk memperoleh kata dasar dilakukan proses *string similarity* menggunakan metode *n*-gram *stemming*, apabila tingkat kemiripannya memenuhi nilai ambang batas maka kata dasar ditampilkan. Adapun alur proses perancangan sistem dilakukan dalam Gambar 2, Gambar 3, dan Gambar 4.



Gambar 2. Alur Proses Sistem *Stemmer* Bahasa Bali



Gambar 3. *On Page Connector* A *Stemmer* Bahasa Bali



Gambar 4. *Off Page Connector* B *Stemmer* Bahasa Bali

2.2.1. Hapus Prefiks

Bahasa Bali memiliki dua belas bentuk prefiks, berikut merupakan *rules* atau aturan yang diterapkan untuk meluluhkan setiap prefiks pada bahasa Bali.

2.2.1.1. Prefiks *n*

Berikut adalah *rules* untuk proses meluluhkan prefiks *n*:

1. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *ngidih* bentuk dasarnya *idih*.
2. Apabila bentuk dasarnya berawal dengan fonem semivokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *ngwangun* bentuk dasarnya *wangun*.
3. Apabila bentuk dasarnya berawal dengan fonem konsonan *t*, *d*, maka dibubuhkan alomorf *n*, untuk memperoleh bentuk dasar dilakukan pengubahan fonem *n* menjadi *t* atau *d*, contoh: *negul* bentuk dasarnya *tegul* dan *nundun* bentuk dasarnya *dundun*.
4. Apabila bentuk dasarnya berawal dengan fonem konsonan *c*, *j*, *s*, maka dibubuhkan alomorf *ny*, untuk memperoleh bentuk dasar dilakukan pengubahan fonem *ny* menjadi *c*, *j*, atau *s*, sebagai contoh: *nyacad* bentuk dasarnya *cacad*, *nyaring* bentuk dasarnya *jaring*, dan *nyampat* bentuk dasarnya *sampat*.
5. Apabila bentuk dasarnya berawal dengan fonem konsonan *k*, *g*, maka dibubuhkan alomorf *ng*, untuk memperoleh bentuk dasar dilakukan pengubahan fonem *ng* menjadi *k* atau *g*, sebagai contoh: *ngutang* bentuk dasarnya *kutang* dan *ngambar* bentuk dasarnya *gambar*.
6. Apabila bentuk dasarnya berawal dengan fonem konsonan *p*, *b*, maka dibubuhkan

alomorf *m*, untuk memperoleh bentuk dasar dilakukan pengubahan fonem *m* menjadi *p* atau *b*, contoh: *mapag* bentuk dasarnya *papag* dan *matek* bentuk dasarnya *batek*.

7. Apabila bentuk dasarnya berawal dengan fonem konsonan nasal *m*, *n*, maka dibubuhkan alomorf *nga*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *nga*, contoh: *ngamaling* bentuk dasarnya *maling* dan *nganengneng* bentuk dasarnya *nengneng*.

2.2.1.2. Prefiks *ma*

Berikut adalah *rules* untuk proses meluluhkan prefiks *ma*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *makesiab* bentuk dasarnya *kesiab*.
2. Apabila bentuk dasarnya berawal dengan fonem semivokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *mayasa* bentuk dasarnya *yasa*.
3. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan fonem *a* pada prefiks *ma*, contoh: *mikut* kata dasarnya *ikut* dan *mubad* kata dasarnya *ubad*.

2.2.1.3. Prefiks *pa*

Aturan atau *rules* untuk proses meluluhkan prefiks *pa* dilakukan apabila melekat pada bentuk asal yang dimulai dengan fonem vokal terjadi sandi, dalam hal ini fonem *a* pada prefiks *pa* diluluhkan. Contoh: *pileh* bentuk dasarnya *ileh*.

2.2.1.4. Prefiks *ka*

Aturan atau *rules* untuk proses meluluhkan prefiks *ka* dilakukan apabila melekat pada bentuk asal yang dimulai dengan fonem vokal terjadi sandi, dalam hal ini fonem *a* pada prefiks *ka* diluluhkan. Contoh: *kicen* bentuk dasarnya *icen.*

2.2.1.5. Prefiks *sa*

Berikut adalah *rules* untuk proses meluluhkan prefiks *sa*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *sajagat* bentuk dasarnya *jagat.*
2. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *sausan* bentuk dasarnya *usan*.

2.2.1.6. Prefiks *a*

Berikut adalah *rules* untuk proses meluluhkan prefiks *a*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *adiri* bentuk dasarnya *diri.*
2. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *aukud* bentuk dasarnya *ukud.*

2.2.1.7. Prefiks *pra*

Aturan atau *rules* untuk proses meluluhkan prefiks *pra* dilakukan apabila bentuk dasarnya berawal dengan fonem konsonan, contoh: *prajani* bentuk dasarnya *jani.*

2.2.1.8. Prefiks *pari*

Berikut adalah *rules* untuk proses meluluhkan prefiks *pari*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *paribasa* bentuk dasarnya *basa*.
2. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *pariindik* bentuk dasarnya *indik.*

2.2.1.9. Prefiks *pati*

Aturan atau *rules* untuk proses meluluhkan prefiks *pati* dilakukan apabila bentuk dasarnya berawal dengan fonem konsonan, contoh: *patigrape* bentuk dasarnya *grape*.

2.2.1.10. Prefiks *maka*

Berikut adalah *rules* untuk proses meluluhkan prefiks *maka*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *makasami* bentuk dasarnya *sami*.
2. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *makaukud* bentuk dasarnya *ukud.*

2.2.1.11. Prefiks *saka*

Berikut adalah *rules* untuk proses meluluhkan prefiks *saka*:

1. Apabila bentuk dasarnya berawal dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *sakabesik* bentuk dasarnya *besik*.
2. Apabila bentuk dasarnya berawal dengan fonem vokal, maka untuk memperoleh bentuk dasar dilakukan penghapusan prefiks, contoh: *sakaukud* bentuk dasarnya *ukud*.

2.2.1.12. Prefiks *kuma*

Aturan atau *rules* untuk proses meluluhkan prefiks *kuma* dilakukan apabila bentuk dasarnya berawal dengan fonem konsonan, contoh: *kumajaum* bentuk dasarnya *jaum*.

2.2.2. Hapus Sufiks

Bahasa Bali memiliki delapan bentuk sufiks, berikut merupakan *rules* atau aturan yang diterapkan untuk meluluhkan setiap sufiks pada bahasa Bali.

2.2.2.1. Sufiks *a*

Berikut adalah *rules* untuk proses meluluhkan sufiks *a*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *daara* bentuk dasarnya *daar*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf na, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *na* pada sufiks *a*, contoh: *anggona* bentuk dasarnya *anggo*.

2.2.2.2. Sufiks *ang*

Berikut adalah *rules* untuk proses meluluhkan sufiks *ang*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *jemakang* bentuk dasarnya *jemak*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf *nang* atau *yang*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *nang* atau *yang* pada sufiks *ang*, contoh: *gedenang* atau *gedeyang* bentuk dasarnya *gede*.

2.2.2.3. Sufiks *an*

Berikut adalah *rules* untuk proses meluluhkan sufiks *an*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *cenikan* bentuk dasarnya *cenik*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf *nan*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *nan* pada sufiks *an*, sebagai contoh: *dawanan* bentuk dasarnya *dawa*.

2.2.2.4. Sufiks *in*

Berikut adalah *rules* untuk proses meluluhkan sufiks *in*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *jagurin* bentuk dasarnya *jagur*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf *nin*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *nin* pada sufiks *in*, sebagai contoh: *jumunin* bentuk dasarnya *jumu*.

2.2.2.5. Sufiks *e*

Berikut adalah *rules* untuk proses meluluhkan sufiks *e*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *payuke* bentuk dasarnya *payuk*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf *ne*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *ne* pada sufiks *e*, contoh: *bajune* bentuk dasarnya *baju*.

2.2.2.6. Sufiks *ne*

Berikut adalah *rules* untuk proses meluluhkan sufiks *ne*:

1. Apabila bentuk dasarnya berakhiran dengan fonem konsonan, maka untuk memperoleh bentuk dasar dilakukan penghapusan sufiks, contoh: *baasne* bentuk dasarnya *baas*.
2. Apabila bentuk dasarnya berakhiran dengan fonem vokal, maka dibubuhkan alomorf *nne*, untuk memperoleh bentuk dasar dilakukan penghapusan fonem *nne* pada sufiks *ne*, sebagai contoh: *giginne* bentuk dasarnya *gigi*.

2.2.2.7. Sufiks *n*

Aturan atau *rules* untuk proses meluluhkan sufiks *n* dilakukan apabila bentuk dasarnya berakhiran dengan fonem vokal, contoh: *bukun* bentuk dasarnya *buku*.

2.2.2.8. Sufiks *ing*

Aturan atau *rules* untuk proses meluluhkan sufiks *ing* dilakukan apabila bentuk dasarnya berakhiran dengan fonem vokal, contoh: *rikalaning* bentuk dasarnya *rikala*.

2.2.3. Hapus Infiks

Bahasa Bali memiliki empat bentuk infiks, berikut merupakan *rules* atau aturan yang diterapkan untuk meluluhkan setiap infiks pada bahasa Bali.

2.2.3.1. Infiks *in*

Berikut adalah *rules* untuk proses meluluhkan infiks *in*:

1. Apabila bentuk dasarnya diawali dengan fonem konsonan, maka infiks *in* disisipkan pada suku awal bentuk dasarnya diantara huruf konsonan pertama dan vokal yang mengikutinya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, contoh: *sinurat* bentuk dasarnya *surat*.
2. Apabila bentuk dasarnya diawali dengan fonem vokal, maka infiks *in* disisipkan di awal bentuk dasarnya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, sebagai contoh: *inucap* bentuk dasarnya *ucap*.

2.2.3.2. Infiks *um*

Berikut adalah *rules* untuk proses meluluhkan infiks *um*:

1. Apabila bentuk dasarnya diawali dengan fonem konsonan, maka infiks *um* disisipkan pada suku awal bentuk dasarnya diantara huruf konsonan pertama dan vokal yang mengikutinya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, contoh: *rumaksa* bentuk dasarnya *raksa*.
2. Apabila bentuk dasarnya diawali dengan fonem vokal, maka infiks *um* disisipkan di awal bentuk dasarnya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, sebagai contoh: *umawak* bentuk dasarnya *awak*.

2.2.3.3. Infiks *el*

Aturan atau *rules* untuk proses meluluhkan infiks *el* dilakukan apabila bentuk dasarnya diawali dengan fonem konsonan, maka infiks *el* disisipkan pada suku awal bentuk dasarnya diantara huruf konsonan pertama dan vokal yang mengikutinya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, contoh: *telapak* bentuk dasarnya *tapak*.

2.2.3.4. Infiks *er*

Aturan atau *rules* untuk proses meluluhkan infiks *er* dilakukan apabila bentuk dasarnya diawali dengan fonem konsonan, maka infiks *er* disisipkan pada suku awal bentuk dasarnya diantara huruf konsonan pertama dan vokal yang mengikutinya, untuk memperoleh bentuk dasar dilakukan penghapusan infiks, contoh: *gerudug* bentuk dasarnya *gudug*.

2.2.4. Hapus Konfiks

Bahasa Bali memiliki empat bentuk konfiks, berikut merupakan *rules* atau aturan yang diterapkan untuk meluluhkan setiap konfiks pada bahasa Bali.

2.2.4.1. Konfiks *pa - an*

Konfiks *pa - an* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *pa* dan sufiks *an*, contoh: *pasirepan* bentuk dasarnya *sirep*.

2.2.4.2. Konfiks *ka - an*

Konfiks *ka - an* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ka* dan sufiks *an*, contoh: *kasengsaraan* bentuk dasarnya *sengsara*.

2.2.4.3. Konfiks *ma - an*

Konfiks *ma - an* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ma* dan sufiks *an*, contoh: *majemakan* bentuk dasarnya *jemak*.

2.2.4.4. Konfiks *bra - an*

Konfiks *bra - an* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *bra* dan sufiks *an*, contoh: *bramahan* bentuk dasarnya *amah*.

2.2.5. Hapus Simulfiks

Bahasa Bali memiliki dua bentuk simulfiks, berikut merupakan *rules* atau aturan yang diterapkan untuk meluluhkan setiap simulfiks pada bahasa Bali:

2.2.5.1. Simulfiks *ma - n*

Simulfiks *ma - n* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ma* dan diikuti prefiks *n*, contoh: *mamuduh* bentuk dasarnya *buduh*.

2.2.5.2. Simulfiks *pa - n*

Simulfiks *pa - n* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *pa* dan diikuti prefiks *n*, contoh: *pangalung* bentuk dasarnya *kalung*.

2.2.6. Hapus Kombinasi Afiks

Bahasa Bali memiliki tiga bentuk kombinasi afiks, berikut merupakan *rules* yang diterapkan

untuk meluluhkan setiap kombinasi afiks pada bahasa Bali.

2.2.6.1. Kombinasi Afiks *ma - an*

Kombinasi afiks *ma - an* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ma* dan diikuti sufiks *an*, sebagai contoh: *makurenan* bentuk dasarnya *kuren*.

2.2.6.2. Kombinasi Afiks *ma - n - in*

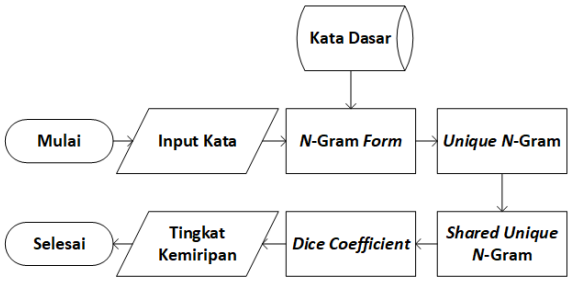
Kombinasi afiks *ma - n - in* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ma*, prefiks *n*, dan diikuti sufiks *in*, contoh: *manuturin* bentuk dasarnya *tutur*.

2.2.6.3. Kombinasi Afiks *ma - n - ang*

Kombinasi afiks *ma - n - ang* dalam hal melekat pada bentuk dasar mengikuti kaidah prefiks *ma*, prefiks *n*, dan sufiks *ang*, contoh: *mangorahang* bentuk dasarnya *orah*.

2.2.7. *N*- Gram *Stemming*

Apabila *rules* yang tersedia tidak dapat mengenali kata berafiks, maka dilakukan proses *string similarity* menggunakan metode *n*-gram *stemming*. Karakteristik kata berafiks yang tidak dapat dikenali oleh *rules* adalah adanya kesalahan dalam menuliskan kata berafiks, seperti kata *maajalan*, *mejalan*, *mjalan*. Adapun alur proses metode *n*-gram *stemming* dilakukan pada Gambar 5.



Gambar 5. Alur Proses *N*-Gram *Stemming*

Pada langkah pertama kata berafiks dan kata dasar pada *database* diubah ke bentuk *n*-gram kemudian dibandingkan dengan cara menghitung jumlah *unique n*-gram atau jumlah karakter *n*-gram yang dihasilkan dan *shared unique n*-gram atau jumlah karakter *n*-gram yang sama antara kata berafiks dan kata dasarnya. Tingkat kemiripan antara *n*-gram kata berafiks dan *n*-gram kata dasar diukur menggunakan metode *dice coefficient* mengikuti persamaan (1).

$$dc = (2 \cdot c) \div (a + b) \quad (1)$$

Pada persamaan (1), *dc* merupakan *dice coefficient*. *c* merupakan *shared unique n*-gram antara dua kata, sedangkan *a* dan *b* merupakan *unique n*-gram. Adapun tahap mengukur tingkat kemiripan antara dua kata *majalan* dan *jalan* menggunakan *n*-gram *stemming*, dengan jumlah karakter *n* dua atau *bi*-gram dan tiga atau *tri*-gram dilakukan dalam Tabel 1.

2.2.8. Menentukan Nilai Ambang Batas

Ketika hasil perhitungan *n*-gram *stemming* antara kata input dan kata dasar memenuhi nilai ambang batas maka kata dasar ditampilkan. Kisaran nilai ambang batas yang digunakan adalah 0,70, 0,65 (Sembok & Bakar, 2011), dan 0,60 (Adamson & Boreham, 1974).

2.2.9. Menghitung Akurasi *Stemming*

Untuk menghitung hasil akurasi *stemming* kami menggunakan persamaan (2) (Husain, 2012).

$$s = \frac{t}{n} \times 100 \quad (2)$$

Pada persamaan (2), *s* merupakan akurasi *stemming*. *t* merupakan jumlah kata yang di *stemming* dengan benar. *n* merupakan jumlah kata berafiks.

3. HASIL DAN PEMBAHASAN

Untuk membuktikan metode yang diusulkan dapat bekerja secara optimal, kami melakukan serangkaian pengujian, seperti membandingkan hasil *stemming* metode yang diusulkan dengan metode terdahulu, menentukan nilai ambang batas yang paling sesuai pada seluruh variasi afiks, dan membandingkan beberapa variasi karakter *n*-gram pada metode *n*-gram *stemming*.

Implementasi serta pengujian sistem dilakukan pada lingkungan pengembangan perangkat lunak sebagai berikut: OS Windows 10 Education 64-*bit*, Prosesor AMD A12-9720P, RAM 8,00GB, IDE Microsoft Visual Studio Enterprise 2017, Bahasa PYTHON 3.6 dengan *Packages* NLTK 3.3.

Data yang digunakan adalah 1000 *list* kata dasar yang dikumpulkan dari sejumlah dokumen buku dan artikel berita *online* berbahasa Bali. Sedangkan pada tahap pengujian kami menggunakan 15 *query*. Adapun kelima belas *query* yang digunakan terdapat pada Tabel 2, dimana pada baris kesebelas hingga kelima belas merupakan *query* yang digunakan untuk menguji metode *n*-gram *stemming*, dimana terdapat beberapa kata berafiks yang tidak dikenali oleh sistem *rule-based*, seperti kata berafiks yang ditulis salah serta adanya kata yang mengalami proses disimilasi.

Hal yang melatarbelakangi penggunaan metode *n*-gram *stemming* adalah kata yang memiliki kesamaan struktural lebih tinggi cenderung sama dengan artinya (Adamson & Boreham, 1974), (Sembok & Bakar, 2011). Namun akurasi *stemming* menggunakan pendekatan statistik tergantung

dengan data *training* yang digunakan (Alotaibi & Gupta, 2018).

Tabel 1. Mengukur Kemiripan Kata dengan *N*-Gram *Stemming*

	***Bi*-Gram**	***Tri*-Gram**
majalan	**m, ma, aj, ja, al, la, an, n**	***m, *ma, maj, aja, jal, ala, lan, an*, n***
jalan	**j, ja, al, la, an, n**	***j, *ja, jal, ala, lan, an*, n***
a	8	9
b	6	7
c	5	5
dc	0,71	0,62

Tabel 2. Daftar *Query* Pengujian

No.	*Query*	Deskripsi
1	*i meme ngajak i bape negakin sepeda.*	prefiks: *ngajak*. konfiks: *negakin*.
2	*semengan kuluke ngongkong.*	sufiks: *semengan, kuluke*. prefiks: *ngongkong*.
3	*palajahin made nganti mamuduh teken i luh.*	prefiks yang mengalami proses disimilasi: *palajahin*. prefiks: *nganti*. simulfiks: *mamuduh*.
4	*telapak liman made beseh ulian dibi majaguran.*	infiks: *telapak*. sufiks: *liman, ulian*. konfiks: *majaguran*.
5	*made lan i luh makurenan duang dasa tiban.*	konfiks: *makurenan*. sufiks: *duang*.
6	*nyen ento menyuling di jabe tengah?*	prefiks: *menyuling*.
7	*sire sane maborbor lulune?*	prefiks: *maborbor*. sufiks: *lulune*.
8	*mangorahang isin hati beline.*	kombinasi afiks: *mangorahang*. sufiks: *isin, beline*.
9	*dadong dauh ngelah siap putih lan sampi aukud.*	prefiks: *ngelah, aukud*.
10	*ngiring lestariang basa baline.*	prefiks: *ngiring*. sufiks: *lestariang, baline*.
11	*tusing nyidangang mekerah ulian meme bapa.*	sufiks yang ditulis salah: *nyidangang*. prefiks yang ditulis salah: *mekerah*. sufiks: *ulian*.
12	*pileh dinene jani made lakar masepedan ke alase.*	prefiks: *pileh*. sufiks yang ditulis salah: *dinene*. konfiks yang ditulis salah: *masepedan*. sufiks: *alase*.
13	*putu melajah mekidung wargasari nganti peteng.*	prefiks yang ditulis salah: *melajah, mekidung*. prefiks: *nganti*.
14	*jero mangku pangelingsir pura tusing nunas pikolih.*	simulfiks yang ditulis salah: *pangelingsir*. prefiks: *nunas*. prefiks yang mengalami proses disimilasi: *pikolih*.
15	*pakeling wenten anak sane melaibang sepeda.*	prefiks yang mengalami proses disimilasi: *pakeling*. konfiks yang ditulis salah: *melaibang*.

Berdasarkan hasil pengujian akurasi *stemming* untuk *query* satu hingga sepuluh yang dihitung menggunakan persamaan (2) terlihat pada Tabel 3, dimana metode yang diusulkan memberikan hasil *stemming* yang lebih baik dibandingkan metode terdahulu, hal ini dikarenakan pada penelitian Nata & Yudiastra hanya meluluhkan dua variasi afiks, yaitu prefiks dan sufiks, kata berafiks seperti *negakin*, *mamuduh*, *majaguran*, dan kata berafiks lain terutama selain prefiks dan sufiks tidak dapat di *stemming* dengan benar.

Sedangkan hasil pengujian untuk *query* sebelas hingga lima belas menggunakan metode *n*-gram *stemming* antara dua variasi karakter *n*-gram, yaitu *bi*-gram dan *tri*-gram dengan kisaran nilai ambang batas 0,6, 0,65, dan 0,70 terlihat pada Tabel 4 hingga Tabel 6 akurasi *stemming* pada setiap kisaran nilai ambang batas, akurasi *bi*-gram lebih baik dibandingkan *tri*-gram, hal ini dikarenakan pada metode *n*-gram *stemming* jumlah *unique n*-gram pada *bi*-gram lebih sedikit dibandingkan *tri*-gram, sedangkan untuk jumlah *shared unique n*-gram pada *bi*-gram dan *tri*-gram bernilai sama, sehingga akurasi *stemming* dengan *bi*-gram memberikan hasil yang lebih baik.

Pada penelitian terdahulu yang dilakukan oleh Adamson & Boreham menyebutkan untuk jumlah *n* dua atau *bi*-gram memberikan beberapa informasi tentang urutan huruf dalam satu kata, sedangkan untuk jumlah *n* tiga atau *tri*-gram memberikan lebih banyak informasi tentang urutan huruf, namun dapat menutupi terjadinya urutan karakter yang lebih pendek yang mungkin signifikan. Selain digunakan untuk memprediksi karakter berikutnya dalam urutan *n* (Putra dkk, 2018), pemanfaatan *n*-gram sering digunakan untuk membangun kamus kata dari data *training* (Abidin & Ferdhiana, 2016), (Mathew & Bai, 2016).

Akurasi *stemming* pada setiap kisaran nilai ambang batas pada *query* kesebelas hingga kelima belas memperlihatkan semakin tinggi kisaran nilai ambang batas maka akurasi *stemming* akan semakin rendah, hal ini dikarenakan perbedaan urutan huruf antara kata yang dibandingkan akan mempengaruhi akurasi *stemming*, sebagai contoh kata berafiks *bukun* dibandingkan dengan kata dasarnya *buku*, dengan *bi*-gram memperoleh hasil 0, 72, padahal perbedaan antara kata *bukun* dan *buku* hanya terletak di satu urutan huruf pada sufiks *n*.

Tabel 3. Hasil Pengujian *Query* 1 - 10

No.	*Query*	Nata & Yudiastra %	Metode Usulan %
1	Q1	50	100
2	Q2	100	100
3	Q3	33,33	66,67
4	Q4	50	100
5	Q5	50	100
6	Q6	100	100
7	Q7	100	100
8	Q8	66,67	100
9	Q9	100	100
10	Q10	100	100
Average		**75**	**96,67**

Tabel 4. Hasil Pengujian *N*- Gram Stemming *Query* 11 - 15 dengan Nilai Ambang Batasa 0,60

No.	*Query*	Akurasi	
		***bi*-gram %**	***tri*-gram %**
1	Q11	100	33,33
2	Q12	50	50
3	Q13	100	66,67
4	Q14	100	66,67
5	Q15	50	0
Average		**80**	**43,33**

Tabel 5. Hasil Pengujian *N*- Gram Stemming *Query* 11 - 15 dengan Nilai Ambang Batas 0,65

No.	*Query*	Akurasi	
		***bi*-gram %**	***tri*-gram %**
1	Q11	33,33	0
2	Q12	50	25
3	Q13	66,67	66,67
4	Q14	66,67	33,33
5	Q15	50	0
Average		**53,33**	**25**

Tabel 6. Hasil Pengujian *N*- Gram Stemming *Query* 11 - 15 dengan Nilai Ambang Batas 0,70

No.	*Query*	Akurasi	
		***bi*-gram %**	***tri*-gram %**
1	Q11	33,33	0
2	Q12	50	25
3	Q13	66,67	33,33
4	Q14	33,33	33,33
5	Q15	0	0
Average		**36,67**	**18,33**

4. KESIMPULAN

Pada penelitian ini, kami mengombinasikan metode *rule-based* dan *n*-gram *stemming*. Metode *rule-based* digunakan untuk membentuk *rules* yang meluluhkan seluruh variasi afiks sedangkan metode *n*-gram *stemming* digunakan apabila *rules* yang tersedia tidak dapat dikenali. Berdasarkan 1000 *list* kata dasar dan kesepuluh *query* yang diberikan antara metode yang diusulkan dan metode terdahulu, metode yang diusulkan memperoleh akurasi *stemming* lebih baik 96,67% dibandingkan metode Nata & Yudiastra 75%, hal ini dikarenakan *rules* pada metode Nata & Yudiastra hanya meluluhkan dua variasi afiks, yaitu prefiks dan sufiks, sedangkan pada lima *query* yang diberikan untuk *n*-gram *stemming*, kata berafiks yang tidak dikenali oleh *rules* dapat di *stemming* dengan baik, terutama pada *bi*-gram dengan nilai ambang batas 0,60.

Hasil pengujian menunjukan metode yang diusulkan memperoleh akurasi *stemming* lebih baik dari metode Nata & Yudiastra serta metode *n*-gram *stemming* mampu mengenali beberapa kata berafiks yang tidak dapat dikenali oleh *rules*. Pada penelitian berikutnya, kami akan memperhatikan semantik setiap kata dan tahap validasi menggunakan aplikasi *text mining* tentang bahasa Bali (Putra dkk, 2016), yaitu *question answering system*.

DAFTAR PUSTAKA

ABIDIN, T.F. & FERDHIANA, R., 2016. Algorithm for Updating N-Grams Word Dictionary for Web Classification. *International Conference on Informatics and Computing (ICIC)*.

ADAMSON, G.W. & BOREHAM, J., 1974. The Use of An Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. *Information Storage and Retrieval*, vol.10, pp.253–260.

ADRIANI, M., ASIAN, J., NAZIEF, B., TAHAGHOGHI, S.M.M. & WILLIAMS, H.E., 2007. Stemming Indonesian: A Confix Stripping Approach. *ACM Transactions on Asian Language Information Processing (TALIP)*.

ALOTAIBI, F.S. & GUPTA, V., 2018. A Cognitive Inspired Unsupervised Language-Independent Text Stemmer for Information Retrieval. *Cognitive Systems Research*, vol.52, pp.291–300.

BALASANKAR, C., SOBHA, T. & MANUSANKAR, C., 2016. Multi Level Inflection Handling Stemmer using Iterative Suffix Stripping for Malayalam Language. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp.530–534.

DE-ROECK, A.N. & AL-FARES, W., 2000. A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp.199–206.

DINCER, B.T. & KARAOGLAN, B., 2003. Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish. *International Symposium on Computer and Information Sciences (ISCIS)*, pp.244–251.

GRANOKA, I.W.O., NARYANA, I.B.U., JENDERA, I.W., BAWA, I.W., MEDERA, I.N., PUTRAYASA, I.G.N., ANOM, I.G.K., TAMA, I.W., DENES, I.M., PURWA, I.M., SUKAYANA, I.N., & INDRA, I.B.K.M., 1996. Tata Bahasa Baku Bahasa Bali. Balai Penelitian Bahasa Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan, Denpasar.

GROSAN, C. & ABRAHAM, A., 2011. Rule-Based Expert Systems. In: *Intelligent Systems*, Intelligent Systems Reference Library, vol.17, pp.655–696.

HUSAIN, M.S., 2012. An Unsupervised Approach to Develop Stemmer. *International Journal on Natural Language Computing (IJNLC)*,

vol.1, pp.15–23.

KROVETZ, R., 1993. Viewing Morphology as An Inference Process. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.191–202.

LARKEY, L.S., BALLESTEROS, L. & CONNELL, M.E., 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-Occurrence Analysis. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.275–282.

LIGEZA, A., 2006. Logical Foundations for Rule-Based Systems. 2nd edition, Springer, Heildelberg.

LINDSAY, S., 1988. Practical Applications of Expert Systems. John Wiley & Sons Inc., Chichester.

LOVINS, J.B., 1968. Development of A Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, vol.11, pp.22–31.

MAJUMDER, P., MITRA, M., PARUI, S.K., KOLE, G., MITRA, P. & DATTA, K., 2007. YASS: Yet Another Suffix Stripper. *ACM Transactions on Information Systems (TOIS).*

MATHEW, N.V. & BAI, V.R., 2016. Analyzing the Effectiveness of N-Gram Technique Based Feature Set in a Naive Bayesian Spam Filter. *International Conference on Emerging Technological Trends (ICETT).*

MAYFIELD, J. & MCNAMEE, P., 2003. Single N-Gram Stemming. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp.415–416.

MEMET, R., NIJAT, M., MAHMUT, G. & HAMDULLA, A., 2017. A Rule and Statistical Modeling Based Stem Extraction Method for Kazakh Words. *International Conference on Asian Language Processing (IALP)*, pp.231–234.

MOULINIER, I., MCCULLOH, J.A. & LUND, E., 2001. West Group at CLEF 2000: Non-English Monolingual Retrieval. *Cross-Language Information Retrieval and Evaluation (CLEF)*, pp.253–260.

NATA, G.N.. & YUDIASTRA, P.P., 2017. Stemming Teks Sor-Singgih Bahasa Bali. *Konferensi Nasional Sistem dan Informatika*, pp.608–612.

NAZIEF, B. & ADRIANI, M., 1996. Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia. Internal Publication, Faculty of Computer Science, University of Indonesia, Depok.

NIKOLOPOULOS, C., 1997. Expert Systems - Introduction to First and Second Generation and Hybrid Knowledge Based Systems. CRC, Boca Raton.

PAICE, C.D., 1994. An Evaluation Method for Stemming Algorithms. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp.42–50.

PATIL, H.B. & PATIL, A.S., 2017. MarS: A Rule-Based Stemmer for Morphologically Rich Language Marathi. *International Conference on Computer, Communications and Electronics*, pp.580–584.

PORTER, M.F., 1980. An Algorithm for Suffix Stripping. *Program*, vol.14, pp.130–137.

PORTER, M.F., 2001. Snowball: A Language for Stemming Algorithms.

PRAMUDITA, Y.D., PUTRO, S.S. & MAKHMUD, N., 2018. Klasifikasi Berita Olahraga menggunakan Metode Naive Bayes dengan Enhanced Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol.5, no.3, pp.269–276.

PUTRA, I.B.G.W., SUDARMA, M. & KUMARA, I.N.S., 2016. Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naive Bayes Classifier. *Teknologi Elektro*, vol.15, no.2, pp.81–86.

PUTRA, S.J., GUNAWAN, M.N. & SURYATNO, A., 2018. Tokenization and N-Gram for Indexing Indonesian Translation of the Quran. *International Conference on Information and Communication Technology (ICoICT).*

SEMBOK, T.M. & BAKAR, Z.A., 2011. Effectiveness of Stemming and N-Grams String Similarity Matching on Malay Documents. *International Journal of Applied Mathematics and Informatics*, vol.5, pp.208–215.

TALA, F.Z., 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.