HASIL CEK_NLP; Kesalahan Penulisan; Koreksi Kata; Pengoreksi Kesalahan Kata; Pemrosesan Bahasa Alami

by Daffa Setiawan Suparno, Miftahurrahma Rosyda Penggunaan Text Modeling Untuk Identifikasi Kesala

Submission date: 13-Aug-2022 01:23PM (UTC+0700)
Submission ID: 1881995482
File name: mas_daffa.docx (187.56K)
Word count: 4466
Character count: 27627

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

Penggunaan *Text Modeling* Untuk Identifikasi Kesalahan Penulisan Kata Pada Teks Pidato Bupati Banggai Sulawesi Tengah

Daffa Setiawan Suparno[1], Miftahurrahma Rosyda[2,*]

Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: [1]Daffa1700018027@webmail.uad.ac.id, [2,*]miftahurrahma.rosyda@tif.uad.ac.id
Email Penulis Korespondensi: miftahurrahma.rosyda@tif.uad.ac.id

Abstrak−Kesalahan pengetikan atau tipografi merupakan sebuah kesalahan yang dibuat saat melakukan pengetikan suatu dokumen atau teks, kesalahan pengetikan bisa terjadi karena kegagalan mekanis atau slip tangan maupun jari. Umumnya kesalahan pengetikan merupakan suatu yang sering terjadi pada saat seseorang sedang mengetik dan dianggap lumrah, namun kesalahan pengetikan ini dalam beberapa konteks dapat mengubah arti kata atau bakan arti dari kalimat itu sendiri. Hal ini yang menyebabkan diperlukannya pengoreksian Kembali setelah seseorang telah selesai mengetik. Tetapi proses pengoreksian kebanyakan masih secara manual sehingga hasil dari koreksi tergantung dari seberapa teliti seseorang melakukan pengoreksian dan seberapa banyak dokumen yang akan dikoreksi. Oleh karena itu diperlukan sebuah sistem yang dapat melakukan pengoreksian secara cepat dan akurat, proses pengoreksian dapat dilakukan dengan berbagai metode salah satunya menggunakan metode text modelling. Pada penelitian ini data uji menggunakan dokumen pidato penting Bupati Banggai Sulawesi Tengah sebanyak 10 dokumen. Metode text modelling dapat dokombinasikan dengan metode pendukung lainnya seperti word2vec, dimana word2vec akan digunakan sebagai rekomendasi kata hasil koreksi. Penelitian ini membuat sebuah sistem yang dapat mengoreksi kesalahan kata pada dokumen pidato penting Bupati Banggai Sulawesi Tengah dengan menggunakan metode text modelling serta Word2Vec, hasil yang didaptkan dari sistem yang telah dibuat adalah sistem memiliki kinerja yang baik serta mendaptkan hasil pengujian yang maksimal.

Kata Kunci: NLP; Kesalahan Penulisan; Koreksi Kata; Pengoreksi Kesalahan Kata; Pemrosesan Bahasa Alami

Abstract−Typing errors or typography are errors made when typing a document or text, typing errors can occur due to mechanical failure or slipping of the hand or finger. Generally, typing errors are something that often occurs when someone is typing and is considered normal, but this typing error in some contexts can change the meaning of the word or even the meaning of the sentence itself, This causes the need for correction again after someone has finished typing, but the correction process is mostly still manually so the results of the correction depend on how carefully someone makes corrections and how many documents will be corrected. Therefore we need a system that can make corrections quickly and accurately, the correction process can be done by various methods, one of which is using the text modeling method. In this study, the test data used 10 documents of the Banggai Regent's important speech, Central Sulawesi. The text modeling method can be combined with other supporting methods such as word2vec, where word2vec will be used as a recommendation for corrected words. This study creates a system that can correct word errors in important speech documents of the Banggai Regent, Central Sulawesi by using text modeling and Word2Vec methods, the results obtained from the system that has been made are the system has good performance and gets maximum test results.

Keywords: NLP; Writing Mistake; Typo Correction; Writing Correction; Natural Language Processing

1. PENDAHULUAN

Pidato merupakan kegiatan berbicara yang disampaikan dengan susunan yang baik kepada orang banyak yang berupa komunikasi yang digunakan dalam forum resmi, berpidato juga merupakan suatu wujud kegiatan berbahasa lisan yang mementingkan ekspresi gagasan dan penalaran yang didukung oleh aspek nonkebahasaan[1]. Dalam penyampaiannya sebuah pidato mempunyai krakteristik sendiri yaitu, pidato harus menggunakan bahasa yang mudah dimengerti dan dipahami oleh para pendengarnya serta isi dari pidato harus jelas, objektif, mengandung kebenaran dan tidak menimbulkan pertentangan[2]. berdasarkan karakteristik dari sebuah pidato yang disebutkan diatas, tidak menutup kemungkinan pidato yang dibuat dapat terjadi kesalahan pengetikan atau tipografi, kesalahan pengetikan dapat dikategorikan menjadi dua macam yaitu disengaja maupun tidak disengaja.

Tipografi adalah kesalahan pengetikan yang dapat desebabkan oleh kegagalan mekanis seorang penulis[3], ada beberapa faktor yang dapat mempengaruhi dalam pembuatan sebuah teks pidato diataranya a) Terjadi slip jari, b) Ketidak telitian seorang penulis, c) Tidak fokus dalam menulis, d) Kurangnya pengetahuan penulis dalam menggunakan bahasa baku, dan d) Kebiasaan penulis menggunakan bahasa sehari-hari.

Beberapa kendala tersebut dapat berpengaruh penting bagi tingkat kualitas pidato yang akan dibuat. Salah satu cara untuk menghindari kesalahan pengetikan ini adalah melakukan pengoreksian secara manual, namun kemampuan seseorang melakukan pengoreksian secara manual tergantung dari tingkat ketelitian mereka dan seberapa banyak teks yang akan dikoreksi, oleh karena itu diperlukan sebuah sistem yang dapat mengoreksi kesalahan kata secara cepat dan akurat dengan tujuan untuk membantu user dalam proses pengoreksian dokumen.

Dataset merupakan bagian terpenting yang harus disiapkan dalam penelitian ini dataset pada penelitian ini menggunakan korpus dan Word2Vec yang berisikan bentuk kata dasar berbahasa Indonesia, dimana dataset korpus digunakan sebagai koreksi kata pembanding untuk mengambil keputusan koreksi kata yang salah. Dan dataset Word2Vec digunakan sebagai rekomendasi kata setelah kata tersebut dikoreksi.

Algoritma Nazief & Adriani merupakan algoritma yang dikembangkan oleh Bobby Nazief dan Mirna Adriani, algoritma ini merupakan bagian dari text modelling. Algoritma nazief adriani memiliki kemampuan untuk

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

melakukan stemming lebih akurat pada dokumen berbahasa Indonesia dibandingkan dengan lagoritma sejenis, hal ini dikarenakan pada algoritma Nazief dan Adriani terdapat beberapa penambahan aturan untuk reduplikasi, serta penambahan aturan untuk awalan dan akhiran yang bertujuan untuk meningkatkan presisi dari setiap kata[4], sehingga algoritma ini sangat cocok diterapkan untuk penelitian kali ini.

Word2Vec merupakan sebuah metode dalam ruang vektor yang sangat efektif digunakan untuk mempelajari dan merepresentasi ruang vektor dari suatu kata dengan cara pemetaan. Word2Vec digunakan sebagai penghubung antara kata yang memiliki maksud atau makna sejenis[5]. Pada penelitian terdahulu yang dilakukan oleh Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avianash Balakrishnan, Pin-Yu Chen, Pradeep Rvikumar, Michael J. Witbrock dengan judul "Word Mover;s Embedding : From Word2Vec to Document Embedding" didapatkan kesimpulan bahwa Word2Vec sangat disarankan untuk digunakan dalam penelitian pada dokumen yang memiliki banyak teks[6].

Pada penelitian sebelumnya yang dilakukan oleh Manase Sahat H Simarangkir dengan judul penelitian "Studi Perbandingan Algoritma – Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia" menggunakan metode stemming vega dan stemming Nazief & Adriani hasil yang didapatkan dari perbandingan dua algoritma tersebut adalah algoritma Nazief Adriani merupakan algoritma terbaik dikarenakan mendapatkan nilai akurasi yang paling tinggi yaitu 97,93% dibandingkan menggunakan algoritma vega dengan tingkat akurasi 63,48%[4].

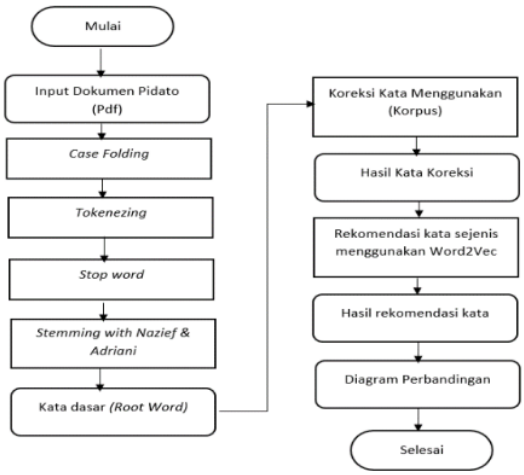
Penelitian lain yang dilakukan oleh Adhi Prasidhatama dan Kristien Margi Suryaningrum dengan judul "Perbandingan Algoritma Nazief & Adriani dengan Algoritma Idris Untuk Pencarian Kata Dasar" menggunakan metode Nazief Adriani dan algoritma Idris, mendapatkan hasil algoritma stemming idris memiliki waktu yang lebih cepat dengan waktu paling cepat 8.15 detik dibandingkan algoritma stemming Nazief & Adriani dengan waktu 10.27 detik, namun algoritma Nazief & Adriani memiliki tingkat akurasi yang lebih baik yaitu sebesar 99.68% dibandingkan dengan algoritma Idris sebesar 92.84% [7].

Berdasarkan penjelasan yang telah dijabarkan diatas maka penelitian ini akan membuat sebuah sistem yang dapat mengoreksi, dan memberikan rekomendasi kata yeng telah dikoreksi, dengan menggunakan metode berbeda yaitu menggunakan Algoritma Nazief Adrini dan Word2Vec sebagai rekomendasi kata yang akan dibandingkan, serta akan melakukan pengujian menggunakan Confusion Matrix.

2. METODOLOGI PENELITIAN

2.1 Metodologi Penelitian

Metodologi penelitian bersifat penting bagi sebuah penelitian, metodologi penelitian merupakan tahapan tahapan secara sistematis yang dilakukan pada sebuah penelitian sehingga penelitian dapat terarah dengan baik. Dapat terlihat pada gambar dibawah ini.



Gambar 1. Flowchart Perancangan Sistem Penelitian

2.2 Subyek Penelitian

Subyek penelitian berlokasi di kantor Bupati Banggai, Sulawesi Tengah, Gedung Humas dan Protokoler Kabupaten Banggai, sampel yang diambil dalam penelitian ini adalah sampel berupa arsip pidato Bupati Banggai sejak tahun 2019-2020 sebanyak 52 sampel pidato. Sampel pidato yang diambil nantinya akan dijadikan sebagai data uji dalam penelitian kali ini.

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

2.2 Motode Yang Digunakan

Metode metupakan hal penting dalam sebuah penelitian baru, kebaruan suatu metode sangat berdampak penting bagi dunia ilmu pengetahuan, dengan adanya suatu metode maka penelitian tersebut akan mendapatkan hasil yang maksimal. Untuk lebih jelasnya dapat dilihat pada penjelasan dibawah ini mengenai metode yang akan digunakan pada penelitian kali ini:

a. Nazief & Adriani
   Algoritma Nazief & Adriani merupakan suatu algoritma yang berdasar pada aturan morfologi bahasa Indonesia yang mengelompokan imbuhan, yaitu imbuhan yang diperbolehkan *(allowed affixes)* dan imbuhan yang tidak diperbolehkan *(disallowed affixes)*. Pengelompokan tersebut termasuk imbuhan depan (awalan), imbuhan tengah (sisipan), imbuhan belakang (akhiran), serta kombinasi keduanya imbuhan awalan dan akhiran (konfiks).[8] algoritma Nazief & Adriani menggunakan kamus kata dasar dan mendukung adanya *recording*, *recording* merupakan tahapan penyusunan Kembali kata-kata yang telah mengalami proses *stemming* secara berlebihan[9]. Dengan kemampuan tersebut membuat algoritma Nazief Adriani sangat cocok digunakan untuk melakukan proses *Stemming* pada dokumen berbahasa Indonesia.

b. *Word2Vec*
   Word2Vec merupakan suatu metode yang digunakan untuk merepresentasikan setiap kata dalam konteks sebagai vektor dengan N dimensi. Dalam prosesnya *Word2Vec* menerapkan *neural network* yang digunakan untuk menghitung *contextual and semantic similarity* (kesamaan kontekstual dan semantic) dari setiap kata yang dimasukan[10]. Hasil dari *Contextual semantic* tersebut dapat direpresentasikan untuk mendapatkan relasi suatau kata dengan kata lainnya yang memiliki makna yang sama. Sebagai contoh yaitu 'Yogyakarta-Indonesia', 'Pria-Wanita'. Dengan demikian maka jika nantinya Word2Vec diterapkan pada penelitian kali ini diharapkan dapat menemukan kata pengganti (rekomendasi) kata lainnya dari kata yang sudah dikoreksi.

c. *Confusion Matrix*
   Confusion matrix merupakan suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Yang nantinya mendapatkan hasil evaluasi dengan *confusion matrix* sebagai pengujinya yang berupa nilai akurasi, presisi dan *recall*. Presisi dan *recall* merupakan suatau istilah yang akan muncul apabila suatau sistem yang sudah dibuat dapat menampilan hasil dari pengujiannya. Pada *confusion matrix* terdapat empat istilah sebagai representassi hasil klasifikasi. Yaitu *True positive* (TP), *True negative* (TN), *False positive* (FP), dan *False negative* (FN)[11]. Pada penelitian ini menggunakan metode *Confusion Matrix* untuk melakukan pengujian nantinya.

4. HASIL DAN PEMBAHASAN

3.1 Alur Sistem

Pada proses pertama diawali dengan memasukan data uji, dalam hal ini data ujinya adalah teks pidato Bupati banggai, pada proses ini data uji akan diinputkan dengan dokumen berformat pdf. Sehingga setelah data uji dimasukan kedalam sistem nantinya sistem akan langsung melakukan proses *preprocessing*. Dapat dilihat pada gambar dibawah merupakan ilustrasi dari sistem yang akan dibuat.

Tabel 1. Tahapan input text pidato

MENGAWALI SAMBUTAN INI, SAYA MENGAJAK KITA SEMUA UNTUK MEMANJATKAN PUJI DAN SYUKUR ATAS KEHADIRAT ALLAH SWT, TUHAN YANG MAHA ESA, KARENA ATAS RAHMAT DAN KARRUNIANYALA, SEHINGGA PADA HARI INI KITA BERKESEMPATAN UNTUK BERTEMU, MENGHADIRI DAN MENGIKUTI ACARA PEMBUKAAN SOSIALISASI PERATURAN DAERAH

a. Text Preprocessing

Tahap preprocessing adalah proses perubahan bentuk data menjadi lebih terstruktur sesuai kebutuhan untuk proses *text mining* [12]. Text mining merupakan suatu proses otomatis yang dapat membentuk text menjadi lebih terstruktur dan penggalian informasi yang lebih akurat dari sebuah teks. Sehingga dengan kata lain *text preprocessing* merupakan suatu tahapan untuk menyiapkan suatu text mentah menjadi text yang dapat diolah oleh sistem. tahapan-tahapan preprocessing sebagai berikut.

b. Case-Folding

Merupakan tahapan untuk mengubah semuah huruf dalam dokumen menjadi huruf kecil *(lower case)*[12], pada proses ini karakter-karakter 'A'-'Z' yang terdapat pada data akan diubah kedalam karakter 'a'-'z'. sedangkan karakter selain huruf 'a'-'z' (tanda baca dan angka) akan dihilangkan dari data dan dianggap sebagagi delimeter (batas pemisah). Pada tahapan dibawah ini dokumen awal akan melakukan perubahan menjadi huruf kecil secara

Submitted: 03/06/2021; Accepted: 24/06/2021; Published: 31/07/2021

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

keseluruhan. Sehingga pada proses ini huruf kapital akan dirubah menjadi *lower case* dan akan menghapus tanda baca yang tidak diperlukan. berikut contohnya :

Tabel 2. Contoh tahapan *Case Folding*

mengawali sambutan ini saya mengajak kita semua untuk memanjatkan puji dan syukur atas kehadirat allah swt tuhan yang maha esa karena atas rahmat dan karrunianyala sehingga pada hari ini kita berkesempatan untuk bertemu menghadiri dan mengikuti acara pembukaan sosialisasi peraturan daerah

c. Tokenezing

Tokenezing atau proses tokenisasi merupakan suatau proses pemotongan teks menjadi kata, symbol, karakter ataupun tanda baca, sehingga menjadi token yang dapat dianalisa [13]. Namun untuk tanda baca lainnya seperti titik (.), tanda petik (‘), koma (,), dan lain sebagainya akan dianggap sebagai pemisah kata.

Tabel 3. Contoh tahapan *Tokenezing*

- mengawali	- syukur	- rahmat	- menghadiri
- sambutan	- atas	- dan	- dan
- ini	- kehadirat	- karrunianyala	- mengikuti
- saya	- allah	- sehingga	- acara
- mengajak	- swt	- pada	- pembukaan
- kita	- tuhan	- hari	- sosialisasi
- semua	- yang	- ini	- peraturan
- untuk	- maha	- kita	- daerah
- memanjatkan	- esa	- berkesempatan	
- puji	- karena	- untuk	
- dan	- atas	- bertemu	

d. Stopword Removal

Stopword adalah kata kata umum yang sering muncul, tetapi tidak memiliki informasi penting (yang bisa siacuhkan atau dibuang contohnya pembuatan inkes atau daftar kata) contohnya antara lain “yang”, “di”, “ke”, dll. *Stopword removal* adalah rpses *filtering*, atau pemilihan kata-kata penting dari hasil token yang merupakan kata untuk mewakili suatu dokumen [14].

Tabel 4. Contoh tahapan *Stopword Removal*

Sebelum Stopword	Setelah Stopword	Sebelum Stopword	Setelah Stopword
- mengawali	- mengawali	- rahmat	- rahmat
- sambutan	- sambutan	- dan	- karrunianyala
- ini	- mengajak	- karrunianyala	- sehingga
- saya	- semua	- sehingga	- hari
- mengajak	- memanjatkan	- pada	- berkesempatan
- kita	- puji	- hari	- bertemu
- semua		- ini	
- untuk		- kita	
- memanjatkan		- berkesempatan	
- puji		- untuk	
- dan		- bertemu	

e. Stemming

Stemming merupakan proses untuk menghasilkan bentuk kata dsar dari suatu kalimat dengan cara memisahkan kata dari imbuhan kata tersebut baik awalan maupun akhiran[15]. Algoritma Nazief dan Adriani merupakan algoritma yang dikembangkan berdasarkan aturan morfologi bahasa Indonesia yang melakukan pengelompokan imbuhan awalan, akhiran serta sisipan dan menggunakan kamus kata dsar sebagai pencocokan katanya [8]. Dapat dilihat pengaplikasian *Stemming Nazief & Adriani* pada contoh dibawah ini.

Tabel 5. Contoh tahapan *Stemming*

Sebelum Stemming	Setelah Stemming	Sebelum Stemming	Setelah Stemming
- mengawali	- kawal	- rahmat	- rahmat

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

Sebelum Stemming	Setelah Stemming	Sebelum Stemming	Setelah Stemming
- sambutan	- sambut	- karrunianyala	- karrunianyala
- mengajak	- ajak	- sehingga	- hingga
- semua	- semua	- hari	- hari
- memanjatkan	- panjat	- berkesempatan	- sempat
- puji	- puji	- bertemu	- temu

f. Koreksi Menggunakan Kata Dasar Korpus

Korpus *(corpus)* merupakan sekumpulan teks tulisan maupun lisan dalam kehidupan sehari-hari yang jumlahnya sangat banyak yang dapat dijadikan sebagai sumber data penelitian. Korpus biasanya digunakan untuk melakukan penelitian bahasa. Dalam hal ini korpus sendiri memuat unit linguistic (kata, frasa, clasa, kalimat, serta wacana) suatu data dapat dikategorikan sebagai korpus jika kata tersebut berkesinambungan menjadi suatu kesatuan bentuk[16]. Sehingga jika kata tersebut tidak memiliki jumlah kata yang banyak dan berkesinambungan maka kata tersebut tidak bisa dikategorikan sebagai korpus, karena korpus merupakan salah satu data natural dalam melihat pertuturan kondisi masyarakat[17]. Pada koreksi menggunakan kata dasar korpus ini, kata yang telah di stemming akan dilakukan pengoreksian kata sehingga kata yang memiliki ejaan keliru akan disempurnakan dengan mencari kata yang mendekati dengan kata yang akan diuji di dalam dataset korpus.

Tabel 6. Contoh tahapan koreksi menggunakan korpus

Sebelum Koreksi	Setelah Koreksi	Sebelum Koreksi	Setelah Koreksi
- mengawali	- kawal	- rahmat	- rahmat
- sambutan	- sambut	- karrunianyala	- karunianyalah
- mengajak	- ajak	- sehingga	- hingga
- semua	- semua	- hari	- hari
- memanjatkan	- panjat	- berkesempatan	- sempat
- puji	- puji	- bertemu	- temu

g*. Rekomendasi Kata Sejenis Menggunakan *Word2Vec

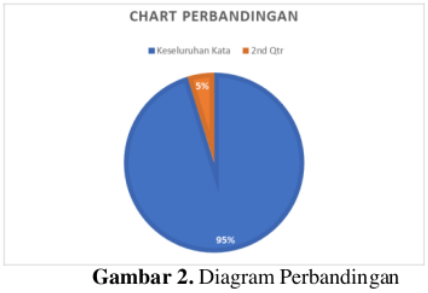
Word2Vec adalah teknik *word embedding* yang dapat mengubah kata menjadi vektor yang terdiri dari sekumpulan angka, kata dari kalimat dapat merepresentasikan makna kata itu sendiri[18]. Cara kerja dari *word2vec* dengan cara mengambil korpus data yang akan digunakan sebagai inputan pembanding. Secara singkatnya word2vec ini merupakan suatu metode yang dapat digunakan untuk menghasilkan word embeddings yang dimana word embeddings sendiri merupakan suatu fitur Natural Language Processing yang setiap katanya memiliki vektor yang sama.

Tabel 7. Rekomendasi Kata Sejenis Menggunakan Word2Vec

Setelah Koreksi	Rekomendasi Kata Sejenis
Karunianyalah	Karunia, Belas kasih, Pemberian, Anugrah.

h. Diagram Perbandingan

Pada tahap terakhir sistem akan menampilkan diagram perbandingan yang berupa sebuah diagram yang di dalamnya memuat tingkatan koreksi kata yang berhasil dikoreksi dan jumlah total kata dalam dokumen.
Contoh:



Gambar 2. Diagram Perbandingan

3.2 Pengujian

Pengujian akurasi sistem ini dilakukan dengan cara melakukan perbandingan antara dokumen sebelum pengujian dengan dokumen setelah pengujian, pada penelitian ini dokumen yang diuji sebanyak 10 dokumen pidato Bupati

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

Banggai Sulawesi Tengah dengan jumlah data sebanyak 7530 kata dengan kesalahan pengetikan sebanyak 60 kata. Selanjutnya dokumen akan diuji menggunakan confusion matrix untuk mendapatkan accuracy, precision, dan recall hasil dari pengujian ditunjukan dalam tabel dibawah ini.

Tabel 8. Pengujian Sistem

Dokumen	Jumlah Kata Di Dalam Dokumen	Jumlah Kata Salah	Kata Salah Sebelum Koreksi	Hasil Koreksi Kata	Rekomendasi Kata Sejenis	Kata Salah (Menurut Ahli)
1	553	7	Mubarokah	Barokah	-	-
			Salawat	Selawat	-	-
			Muslimah	Muslimat	-	-
			Ustadz	Ustaz	-	-
			Shalat	Salat	Sholat Shalat Berjamaah Puasa Ibadah	-
			Irrasional	Irasional	-	-
			Taufiq	Taufik	-	-
2	723	3	Beberpa	Beberapa	Sejumlah Berbagai Dua Banyak Tiga	-
			Taufiq	Taufik	-	-
			Oganisasi	Organisasi	Organisasi Organisasi Lembaga Perkumpulan	-
3	468	4	Assalamuaaikum	Assalamualaikum	-	-
			Iteratif	Interaktif	Multimedia Daring Real-time Grafis Visual	-
			Taufiq	Taufik	-	-
			Wasalamu	Wasalam	-	-
4	782	7	Swuastyastu	Swastiastu	-	-
			Icon	Ikon	Maskot Simbol Landmark Atraksi	-
			Prhatian	Perhatian	Perhatiannya Simpati Sorotan Minat Kritikan	-
			Praktek	Praktik	Cara-cara Metode Etika	-
			Wabillaht	Wabillahi	-	-
			Taufiq	Taufik	-	-
			Wassalamuallaikum	Wassalamualaikum	-	-
5	685	6	Hatur	Matur	-	1
			taufiq	Taufik	-	-
			Sinerjik	Sinergi	Kerjasama Interaksi Hubungan-hubungan Transparansi Kemitraan	-
			Konkrit	Konkret	Logis Objektif	-

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

Dokumen	Jumlah Kata Di Dalam Dokumen	Jumlah Kata Salah	Kata Salah Sebelum Koreksi	Hasil Koreksi Kata	Rekomendasi Kata Sejenis	Kata Salah (Menurut Ahli)
					Menyeluruh	
			Wassalamuallaikum	Wassalamualaikum	-	-
			Ssuai	Sesuai	Disesuaikan Selaras Sejalan Bertentangan Berdasarkan	-
			Swastyastu	Swastiastu	-	-
			Kompentensi	Kompetensi	Profesionalisme Keterampilan Keahlian Mutu	-
6	1011	6	Moderenisasi	Modernisasi	Industrialisasi Pembaharuan Perluasan Ekspansi Restrukturisasi	-
			Management	Manajemen	Pengelolaan Akuntansi Manajerial Pemasaran Penjaminan	-
			Wabillahit	Wabillahi	-	-
			Taufiq	Taufik	-	-
			Suastyastu	Swastiastu	-	-
			Sekretris	Sekretaris	Ketua Sekjen Sekretariat Bendahara Direktur	-
			Aparatr	Aparatur	Pendayagunaan Kepegawaian Agraria Birokrasi Administrasi	-
7	935	8	System	Sistem	Mekanisme Metode Subsistem Struktur	-
			Partisipatif	Antisipatif	-	1
			Perhatianya	Perhatian	Perhatiannya Simpati Sorotan Minat Kritikan	-
			Wabillahit	Wabillahi	-	-
			Shalawat	Selawat	-	
			Lngsung	Langsung	Otomatis Sengaja Sembarangan Spontan	-
			Kata Kata	Katakana	Hiragani Kanji Kana Romanisasi	1
8	725	7	Shalat	Salat	Sholat Shalat Berjamaah Puasa Ibadah	-
			Ridho	Ridha	-	-
			Efesiensi	Efisiensi	Efektivitas Produktivitas	-

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

Dokumen	Jumlah Kata Di Dalam Dokumen	Jumlah Kata Salah	Kata Salah Sebelum Koreksi	Hasil Koreksi Kata	Rekomendasi Kata Sejenis	Kata Salah (Menurut Ahli)
					Keandalan Fleksibilitas	
			Wabarakutuh	Wabarakatuh	-	-
9	730	7	Daearah	Daerah	Wilayah Kawasan Didaerah Kota	-
			Swastyastu	Swastiastu	-	-
			Marjinal	Marginal	Optimal Komunal Homogen Intrinsik	-
			Litrasi	Literasi	-	-
			Kerangka	Rerangka	-	1
			Amiin	Amin	-	-
			Taufiq	Taufik	-	-
10	918	5	Synode	Sinode	Konsili Kongregrasi Gereja GPI	-
			Bathin	Batin	Kebahagiaan Ketenangan Spiritual Kesedihan	-
			Kalang	Alang	Laweh Bunut Lubuk Pauh	1
			Spiritualitas	Spiritualisasi	-	1
			Ngara	Negara	Negaranya Teritori Wilayah Teritorial	-
JUMLAH	7530	60	-	-	-	6

3.1 Pengujian Confusion Matrix

Pengujian akan dilakukan menggunakan metode Confusion matrix, dimana confusion matirix merupakan metode yang sering digunakan untuk melakukan perhitungan akurasi sistem. Confusion matrix juga digunakan utnuk melakukan penganalisaan seberapa baik sistem model klasifikasi yang dibuat dalam mengenali data yang berbeda.

Tabel 9. Pengujian Confusion Matrix

TP	FP	FN	TN
54	6	0	7470

Keterangan:
a. Nilai True Negatif (TN) adalah jumlah data yang dideteksi sebagai negative dan kenyataannya benar.
b. Nilai False Positive (FP) adalah jumlah data positif yang terdeteksi benar.
c. False Negatif (FN) adalah kebalikan dari True Positive.
d. True Positive (TP) merupakan jumlah data dengan kelas positif yang diklasifikasikan sebagai positif.

Dari tabel diatas diketahui bahwa dari 10 dokumen pidato Bupati Banggai dengan jumlah kata sebanyak 7530 terdapat kata yang berhasil dikoreksi sebanyak 60 kata salah, kemudian dari 60 kata tersebut Kembali dikoreksi oleh seorang ahli dari Universitas Negeri Surakarta yaitu Bagus Juniarto Wibowo,S.Pd mahasiswa S2 Pendidikan Bahasa Indonesia dan Sastra. Dan didapatkan hasil koreksi manual terdapat False Positiv (kata yang dianggap benar oleh sistem tetapi salah) sebanyak 6 kata.

$$\text{Akurasi} = \frac{Data\ Yang\ Diprediksi\ Benar\ Dikoreksi + (Jumlah\ Kata\ Berhasil\ Dan\ Tidak\ Berhasil\ Dikoreksi)}{Jumlah\ Data\ Keseluruhan} * 100\ \% \quad (1)$$

$$\text{Akurasi} = \frac{54+7470}{7530} * 100\ \% = 99.92\%$$

$$\text{Presisi} = \frac{Data\ Yang\ Diprediksi\ Benar\ Dikoreksi}{Jumlah\ Data\ Yang\ Dikoreksi} * 100\ \% \quad (2)$$

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

$$\text{Presisi} = \frac{54}{60} * 100\ \% = 90\%$$

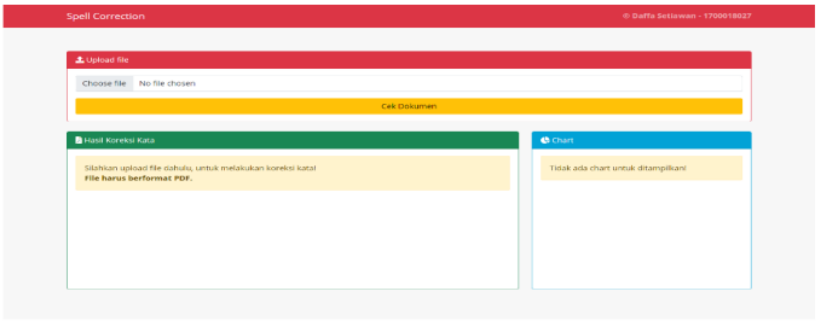
$$\text{Recall} = \frac{Data\ Yang\ Diprediksi\ Benar\ Dikoreksi}{Jumlah\ Keseluruhan\ Data} * 100\ \% \quad (3)$$

$$\text{Recall} = \frac{60}{60} * 100\ \% = 100\%$$

3.2 Implementasi Program

a. Tampilan awal *system*

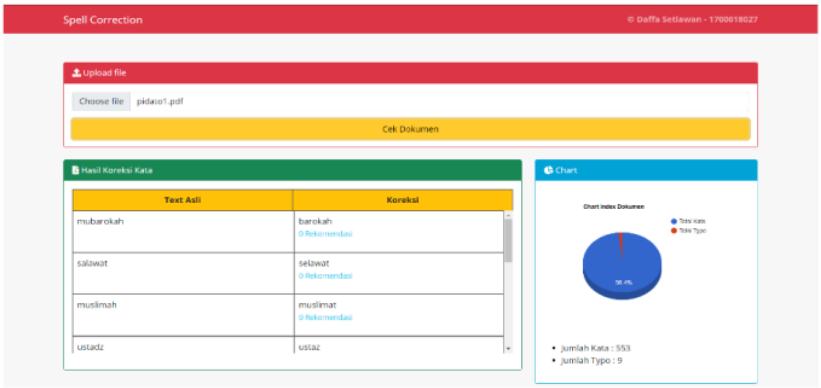
Pada tampilan awal sistem dapat terlihat beberapa menu yaitu menu untuk melakukan upload file dan juga menu eksekusi file dalam bentuk pdf.



Gambar 3. Tampilan Awal Sistem

b. Tampilan Ketika *system* selesai

Pada tampilan ini sistem akan menampilkan teks yang salah *(Typo)* kemudian akan menampilkan koreksi teks yang serta akan menampilkan rekomendasi kata dari kata yang salah tersebut. Setelah sistem menampilkan semuanya maka sistem juga akan menampilkan chart untuk dapat membandingkan seberapa banyak kata yang salah alam dokumen yang telah kita *upload*.



Gambar 4. Tampilan Sistem Koreksi

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dengan judul penggunaan teks modelling untuk identifikasi kesalahan penulisan kata pada teks pidato Bupati Banggai Sulawesi tengah berbasis web, dapat ditarik beberapa kesimpulan sebagai berikut dapat menerapkan dan membuat aplikasi pengoreksi kesalahan kata untuk teks pidato berbahasa Indonesia. Tingkat keakuratan stemming nazief adriani untuk mengoreksi kesalahan kata menggunakan korpus mencapai 92.2% dibandingkan menggunakan metode algoritma *stemming* sejenis. Adapun saran untuk dilakukan pengembangan system kedepannya antara lain, Menambah jumlah dataset system sehingga system dapat memberikan rekomendasi kata yang lebih banyak lagi. Melatih dataset system agar system dapat membedakan kata yang bermakna ganda.

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

UCAPAN TERIMAKASIH

Terima kasih diucapkan kepada Bapak Ir. H. Herwin Yatim, M.M. Bupati Banggai Sulawesi Tengah Periode 2016-2021, yang telah memberikan izin penelitian dan berbagi data untuk penyelesaian penelitian ini. Terima kasih juga diucapkan kepada Univeritas Ahmad Dahlan Yogyakarta yang telah membina serta membimbing saya sehingga penelitian ini dapat terselesaikan. Serta terima kasih yang sebesar besarnya atas bimbingan dosen akademik serta dosen pembimbing saya yang telah meluangkan waktu sehingga penelitian ini dapat terselesaikan.

REFERENCES

[1] A. R. Kusuma, "Penerapan Keterampilan Berbicara Dalam Pidato," 2019.
[2] Susilowati, "Teknik Retorika Dalam Naskah Pidato Nadiem Makarim Pada Hari Guru Nasional 2019," *Trias Polit.*, vol. 4, no. 1, pp. 1–14, 2020.
[3] "Kesalahan tipografi - Wikipedia bahasa Indonesia, ensiklopedia bebas." https://id.wikipedia.org/wiki/Kesalahan_tipografi (accessed Jun. 07, 2021).
[4] M. S. H. Simarangkir, "Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia," *J. Inkofar*, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.
[5] K. W. Church, "Emerging Trends: Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017, doi: 10.1017/S1351324916000334.
[6] L. Wu *et al.*, "Word Mover's Embedding: From Word2Vec to Document Embedding." Accessed: Jun. 07, 2021. [Online]. Available: https://github.
[7] A. Prasidhatama and K. M. Suryaningrum, "Perbandingan Algoritma Nazief & Adriani Dengan Algoritma Idris Untuk Pencarian Kata Dasar," *J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, pp. 1–4, 2018, doi: 10.26905/jtmi.v4i1.1773.
[8] I. P. M. Wirayasa, I. M. A. Wirawan, and I. M. A. Pradnyana, "Algoritma Bastal: Adaptasi Algoritma Nazief & Adriani Untuk Stemming Teks Bahasa Bali," *J. Nas. Pendidik. Tek. Inform.*, vol. 8, no. 1, p. 60, 2019, doi: 10.23887/janapati.v8i1.13500.
[9] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, pp. 49–56, 2017.
[10] "Word2Vec. Word2Vec Model Tutorial (part 1) | by Arif R | Medium." https://arifromadhan19.medium.com/word2vec-95c5df46e045 (accessed Jun. 07, 2021).
[11] H. Judul, "UJI AKURASI APLIKASI AUGMENTED REALITY PEMBELAJARAN HURUF ALFABET BAHASA ISYARAT INDONESIA (BISINDO) PADA VUFORIA MENGGUNAKAN CONFUSION MATRIX."
[12] D. Steveson, H. Agung, and F. Mulia, "Plagiarisme Detection Applications For Tasks and Problems in School Using Rabin Karp Algorithm," Th, May 2018. Accessed: Jun. 07, 2021. [Online]. Available: http://journal.ubm.ac.id/jalu.
[13] "View of ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN." http://ejournal.nusamandiri.ac.id/index.php/inti/article/view/1347/661 (accessed Jun. 07, 2021).
[14] M. S. Anwar, I. M. I. Subroto, and S. Mulyono, "Sistem Pencarian E-Journal Menggunakan Metode Stopword Removal Dan Stemming," *Pros. Konf. Ilm. Mhs. UNISSULA 2*, pp. 58–70, 2019, [Online]. Available: http://lppm-unissula.com/jurnal.unissula.ac.id/index.php/kimueng/article/viewFile/8420/3887.
[15] D. Sebagai *et al.*, "ALGORITMA STEMMING TEKS BAHASA MASSENREMPULU BERBASIS ATURAN TATA BAHASA TUGAS AKHIR."
[16] T. Setiawan, "Korpus dalam kajian penerjemahan," 2017.
[17] B. Bahasa Kalimantan Barat and D. Ari Asfar Balai Bahasa Kalimantan Barat, "CIRI-CIRI BAHASA MELAYU PONTIANAK BERBASIS KORPUS LAGU BALEK KAMPONG CHARACTERISTICS OF PONTIANAK MALAY LANGUAGE BASED ON THE BALEK KAMPONG SONG CORPUS."
[18] J. Nurjaman, R. Ilyas, F. Kasyidi, J. Informatika, U. Jenderal, and A. Yani, "Pengukuran Kesamaan Semantik Pasangan Kalimat Sitasi Menggunakan Convolutional Neural Network," pp. 26–27, 2020.
[1] A. R. Kusuma, "Penerapan Keterampilan Berbicara Dalam Pidato," 2019.
[2] Susilowati, "Teknik Retorika Dalam Naskah Pidato Nadiem Makarim Pada Hari Guru Nasional 2019," *Trias Polit.*, vol. 4, no. 1, pp. 1–14, 2020.
[3] "Kesalahan tipografi - Wikipedia bahasa Indonesia, ensiklopedia bebas." https://id.wikipedia.org/wiki/Kesalahan_tipografi (accessed Jun. 07, 2021).
[4] M. S. H. Simarangkir, "Studi Perbandingan Algoritma - Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia," *J. Inkofar*, vol. 1, no. 1, pp. 40–46, 2017, doi: 10.46846/jurnalinkofar.v1i1.2.
[5] K. W. Church, "Emerging Trends: Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017, doi: 10.1017/S1351324916000334.
[6] L. Wu *et al.*, "Word Mover's Embedding: From Word2Vec to Document Embedding." Accessed: Jun. 07, 2021. [Online]. Available: https://github.
[7] A. Prasidhatama and K. M. Suryaningrum, "Perbandingan Algoritma Nazief & Adriani Dengan Algoritma Idris Untuk Pencarian Kata Dasar," *J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, pp. 1–4, 2018, doi: 10.26905/jtmi.v4i1.1773.
[8] I. P. M. Wirayasa, I. M. A. Wirawan, and I. M. A. Pradnyana, "Algoritma Bastal: Adaptasi Algoritma Nazief & Adriani Untuk Stemming Teks Bahasa Bali," *J. Nas. Pendidik. Tek. Inform.*, vol. 8, no. 1, p. 60, 2019, doi: 10.23887/janapati.v8i1.13500.
[9] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia," *J. Ilm. SINUS*, vol. 15, no. 2, pp. 49–56, 2017.
[10] "Word2Vec. Word2Vec Model Tutorial (part 1) | by Arif R | Medium." https://arifromadhan19.medium.com/word2vec-95c5df46e045 (accessed Jun. 07, 2021).

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 3, Juli 2021, Page 779-789
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i3.3051

[11] H. Judul, "UJI AKURASI APLIKASI AUGMENTED REALITY PEMBELAJARAN HURUF ALFABET BAHASA ISYARAT INDONESIA (BISINDO) PADA VUFORIA MENGGUNAKAN CONFUSION MATRIX."
[12] D. Steveson, H. Agung, and F. Mulia, "Plagiarisme Detection Applications For Tasks and Problems in School Using Rabin Karp Algorithm," Th, May 2018. Accessed: Jun. 07, 2021. [Online]. Available: http://journal.ubm.ac.id/jalu.
[13] "View of ANALISIS SENTIMEN OPINI PUBLIK MENGENAI COVID-19 PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES DAN KNN." http://ejournal.nusamandiri.ac.id/index.php/inti/article/view/1347/661 (accessed Jun. 07, 2021).
[14] M. S. Anwar, I. M. I. Subroto, and S. Mulyono, "Sistem Pencarian E-Journal Menggunakan Metode Stopword Removal Dan Stemming," *Pros. Konf. Ilm. Mhs. UNISSULA 2*, pp. 58–70, 2019, [Online]. Available: http://lppm-unissula.com/jurnal.unissula.ac.id/index.php/kimueng/article/viewFile/8420/3887.
[15] D. Sebagai *et al*., "ALGORITMA STEMMING TEKS BAHASA MASSENREMPULU BERBASIS ATURAN TATA BAHASA TUGAS AKHIR."
[16] T. Setiawan, "Korpus dalam kajian penerjemahan," 2017.
[17] B. Bahasa Kalimantan Barat and D. Ari Asfar Balai Bahasa Kalimantan Barat, "CIRI-CIRI BAHASA MELAYU PONTIANAK BERBASIS KORPUS LAGU BALEK KAMPONG CHARACTERISTICS OF PONTIANAK MALAY LANGUAGE BASED ON THE BALEK KAMPONG SONG CORPUS."
[18] J. Nurjaman, R. Ilyas, F. Kasyidi, J. Informatika, U. Jenderal, and A. Yani, "Pengukuran Kesamaan Semantik Pasangan Kalimat Sitasi Menggunakan Convolutional Neural Network," pp. 26–27, 2020.

Submitted: **03/06/2021**; Accepted: **24/06/2021**; Published: **31/07/2021**

HASIL CEK_NLP; Kesalahan Penulisan; Koreksi Kata; Pengoreksi Kesalahan Kata; Pemrosesan Bahasa Alami

ORIGINALITY REPORT

4%	2%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mira Veranita, Ramayani Yusuf, Yuda Sahidin, Rini Susilowati, Dian Candra Fatihah, Wiwi Warsiati. "Empowering UMKM Dengan Pemanfaatan Digital Marketing Di Era New Normal (Literasi Media Digital Melalui Webinar)", Jurnal Pengabdian kepada Masyarakat UBJ, 2021 Publication	2%
2	docplayer.info Internet Source	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		