A Swear Word Detection System in SKPD Website Content of Bali Province Government

Ester Melinda[1*], Wayan Gede Suka Parwita[2] iD

[1,2] STMIK STIKOM, Denpasar, Indonesia
*Corresponding author: Melinda909@gmail.com

Abstrak

Website menjadi salah satu sistem layanan informasi yang sangat diperlukan oleh masyrakat untuk melakukan berbagai aktifitas pelayanan publik. Hanya saja pada website perlu dilakukan proses pengecekan kata-kata, untuk menghindari adanya kata umpatan, kususnya pada website layanan pemerintahan. Pengecekan kata umpatan pada website membutuhkan waktu yang lama, yakni dilakukan dengan cara membaca setiap konten website yang akan ditambahkan oleh admin. Sehingga dibutuhkan suatu sistem yang dapat mengecek kata umpatan pada website. Adapun tujuan dari penelitian adalah untuk mempermudah proses pendeteksian kata umpatan pada website SKPD. Dalam penelitian ini, dirancang dan dibangun sistem yang dapat mendeteksi kata umpatan pada saat admin menambahkan dan mengubah konten website serta pada saat admin memvalidasi kritik dan saran yang telah ditambahkan oleh pengunjung dengan cara mengimplementasikan text mining. Penelitian ini menggunakan algoritma Nazief dan Adriani untuk proses stemming serta menggunakan algoritma Term Frequency untuk proses pembobotannya. Metode pengujian yang digunakan adalah accuracy measure. Dari hasil pengujian, didapatkan nilai accuracy sebesar 88%; precision sebesar 78,57; recall sebesar 100%; dan f measure sebesar 88%. Hasil yang diperoleh belum dapat maksimal dikarenakan sistem belum dapat membaca struktur kalimat, sehingga suatu kata yang seharusnya penggunaannya bukan sebagai kata umpatan pada suatu dokumen/teks menjadi terdeteksi oleh sistem sebagai kata umpatan, contohnya kata anjing yang terdeteksi pada teks berita anjing rabies.

Keywords: Pendeteksi Kata Umpatan, Website

Abstract

The website is one of the information service systems needed by the community to carry out various public service activities. It's just that it is necessary to do a word-checking process on the website to avoid swear words, especially on government service websites. Checking swear words on the website takes a long time, done by reading every website content that the admin will add. So we need a system that can check swear words on the website. The purpose of the research is to simplify detecting swear words on the SKPD website. In this study, a system was designed and built that can detect swear words when the admin adds and changes website content and validates criticism and suggestions that visitors have added by implementing text mining. This study uses the Nazief and Adriani algorithms for the stemming process and the Term Frequency algorithm for the weighting process. The test method used is accuracy measure. From the test results, the accuracy value is 88%, precision of 78.57%; recall of 100%; and f measure by 88%. The results obtained have not been maximized because the system has not been able to read the sentence structure, so that a word that should not be used as a swear word in a document/text becomes detected by the system as a swear word, for example, the word dog detected in the text of the news about rabid dogs.

Keywords: Swear Detector, Website

1. INTRODUCTION

The development of information technology makes it easy for the public to obtain information from various regions through various social media sites (Setiawan, 2018; Yoga, 2019). One site that is widely used to find information is website services (Ari & Hanum, 2021). Website is one of the media of information and publications that can present various service information and be accessed anywhere and anytime, without geographical area restrictions. (Hayami et al., 2019; Widagdo et al., 2018). Websites are generally used by various institutions, such as government and schools to provide information about institutions

and services. Through a website page, an institution will be able to present information in the form of text, still or motion pictures, animation, sound, and or a combination of all of them, both static and dynamic, which form a series of interrelated buildings, each of which is connected by networks page (hyperlink) (Putra & Harli, 2021; Sugiartawan et al., 2018). A website designed by an institution will certainly help deliver information to the public so that all administrative processes can be simplified.

One of the iusetutions that uses the website to disseminate information is Lembaga Satuan Kerja Perangkat Daerah (SKPD) Bali Province Government. SKPD of Bali Province is an organization/institution in the local government responsible for implementing deconcentration/government tasks in the Bali Province (Ariana, 2020). SKPD of Bali Province consists of several agencies. Each agency has a website whose content is dynamic so that each agency can manipulate all its content without calling a programmer (Pradipa & Munidewi, 2018; Putri et al., 2017). SKPD of Bali Province Government website content consists of ten categories: news, photos, videos, announcements, vision and mission categories, organizational structure, criticism and suggestions, profiles, sliders, and information. Manipulation of website content from all these categories is carried out by admins in each agency, except for the category of criticism and suggestions added by website visitors.

It's just that there is a problem with the website system of the SKPD of Bali Province Government, namely the amount of time needed to check by reading each website content that the admin will add because most of the website content is in the form of dense content with writing. This check is carried out to avoid any swear words in the content because not a little news or information is taken from outside or news reporters. In addition, when website visitors add criticism and suggestions, it is very likely for visitors to comment with swear words. Website content and comments containing these swear words should not be eligible to be displayed on formal web pages such as this government website. The above problems can occur because there is no feature on the SKPD of the Bali Province Government website that can prevent this.

One of the efforts that can be done is to design a swear word detection system. A swear word detection system is a system that can track swear words on a website (Sahrul & Rahman, 2019; Tjahyanti, 2020). One system that can be used is the UML (Unified Modeling Language) system built using a web programming language with the Laravel framework (Hartina & Masri, 2020). UML (Unified Modeling Language) is a programming language based on graphics and visuals for visualizing, specifying, building, and documenting a software development system based on OO (Object-Oriented) (Mubarak, 2019). UML provides vocabulary and writing in MS Word for communication purposes (Prihandoyo, 2018; Veza & Nurlinda, 2021). A model language is a language that has vocabulary and concepts of writing order/rules and physically represents a system (Henderi, 2018).

Several studies that have been conducted previously revealed that the swear word detection system using the RNN model showed better performance than the ANN model with training, validation, and test accuracy of 94%, 84%, and 84%, respectively (Sahrul & Rahman, 2019). Another study also revealed that in addition to using the RNN model, swear words can also be detected using the SVM, Logistic Regression, and Naive Bayes models, which significantly succeeded in detecting insulting elements in comment texts from social media (Sastrodikoro et al., 2018). Further research revealed that the N-Gram method could detect typing errors in non-standard words in Indonesian written works (Hartina & Masri, 2020). Based on some of the results of these studies, it can be said that detecting swear words and non-standard words can be done using various methods and models. There was no UML (Unified Modeling Language) system in previous studies. It was built using a web programming language with the Laravel framework on the swear word detection system on

the SKPD website content of the Bali Province Government. So that research is focused on this study to facilitate the process of detecting swear words on the SKPD of Bali province website.

2. METHOD

This research was conducted by implementing text mining, where the placement of variables from each formula used in the case study was first carried out. Text mining consists of two main stages, namely preprocessing and processing. The first stage is preprocessing, which consists of several stages, including case folding, which is done by converting all criticisms and suggestions into lowercase letters, tokenizing, breaking sentences into words or phrases, where the results are in the form of data with an array type. This stage also removes all separators such as "period," "comma," and so on by using the str_replace function, filtering is done by eliminating less important words such as the words "yang," "and," "at," and others -Other data on the previous tokenizing results in the form of array data by storing words that are not less important, and the stemming stage is carried out by searching for basic words using the Nazief and Adriani algorithm.

The second stage, the processing stage in this study, was carried out using a word weight calculation algorithm. The algorithm used is the Term Frequency (TF) algorithm. The design of the swear word detection system on the SKPD website content of the Bali Province Government uses the UML (Unified Modeling Language) design. This design uses a use case diagram, namely a swear word detection system diagram. The use case diagram of the swear word detection system can be seen in Figure 1.
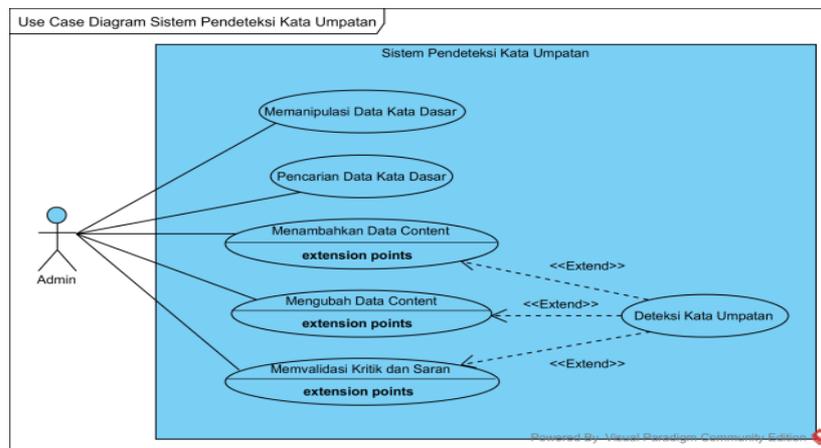

Figure 1. Use Case Diagram of The Swear Word Detection System

3. RESULT AND DISCUSSION

Result

The pages that are used to support the operation of this swear word detection system is the Bali Province Government SKPD admin page, including the basic data master page, the basic word data add form, the basic word data change the form, the basic word data delete page, the content data master page. Form to add data content, form to change data content, and validation page for criticism and suggestions. The system implementation process is carried out in several processes, according to the SKPD admin page. The first process is to provide a basic word data master page used by admins to manipulate basic word data. The basic word data master page display can be seen in Figure 2. The second process is changing

the Swearing Detection Status on the Master Data Content page. Admin can activate or deactivate the swear word detection feature on the master data content page by clicking the on/off detection swear word on the master data content page. The display of the activation of the swear word detection feature on the master data content page can be seen in Figure 3.
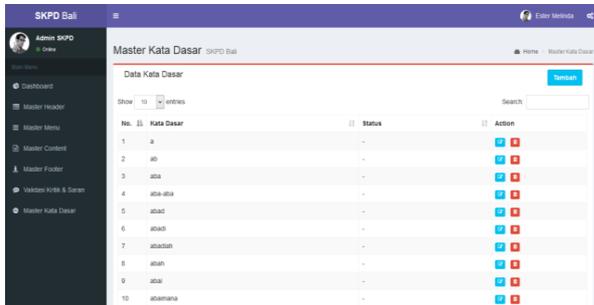

Figure 2. Display of the Master Data page for basic words
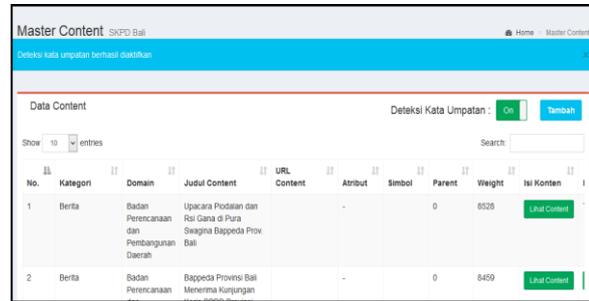
Figure 3. Display of the Activation of the Swear Word Detection Feature on the Master Data Content Page

The third process is the Process of Detecting Swear Words on the Add Data Content Form. The additional data content form is one of the pages where the swear word detection feature is applied. The display of the pop-up form adding data content can be seen in Figure 4.
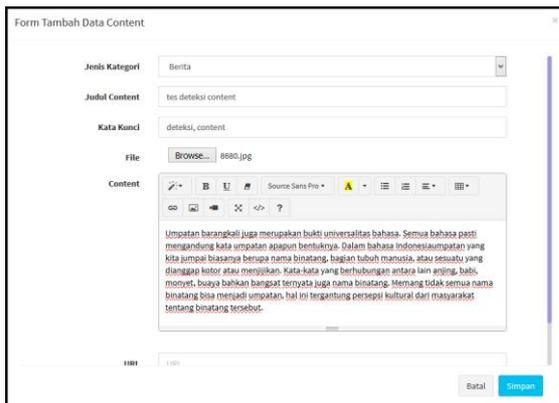
Figure 4. Pop Up Form Display Add Data Content
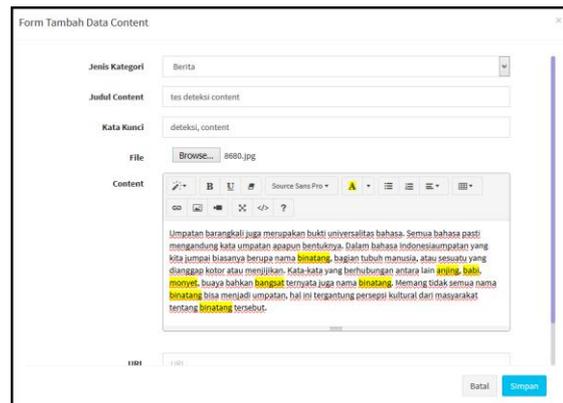
Figure 5. Example of Highlighted Content Display

After the admin types the data contained on the add data content form and presses the save button, the system will detect swear words against the content that the previous admin has typed. The system will display a pop-up confirmation page for adding data content containing the detection results and the message "Is it Do you still want to keep the content?". When the admin wants to cancel the process of storing the content data or want to re-examine the content that contains the swear word, the admin can press the cancel button, so the system will display the pop-up form adding data content again, where the swear words contained in the content have been highlighted. by the system to make it easier for admins to check the content that contains these swear words. An example of the display of highlighted content can be seen in Figure 5. The fourth process is detecting swear words on criticism and suggestions that will be validated. One of the pages where the swear word detection feature is implemented is the criticism and suggestion validation page. The criticism and suggestions validation page display can be seen in Figure 6.
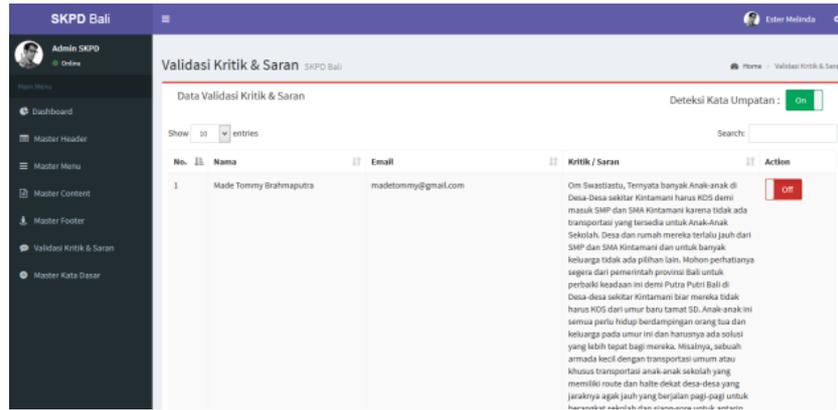
Figure 6. Criticism and Suggestion Validation Page Display

When the admin presses the validation button for criticism and suggestions from off to on, the system will detect swear words against complaints and suggestions that the admin will validate. The system will display a pop-up confirmation page to validate criticism and suggestions containing the detection results and the message "Is it do you still want to display these criticisms and suggestions?". After all the system preparation processes are completed, the system development continues in the system testing process. Using the accuracy measure method, they tested the swear word detection system on SKPD of Bali Province Government website content. This method is used to measure the accuracy of the resulting output, where the values of accuracy, precision, recall, and f measure are carried out. The page used for testing on this system is only the additional data content form because the detection on the change data content form and the criticism and suggestion validation page use the same process, namely by calling the swear word detection function and sending parameters in the form of strings, where the contents of these parameters will be detected swear words. The data tested are in the form of eleven news content texts containing swear words and fourteen news content texts that do not contain swear words to test the accuracy of the output produced by the swear word detection system. The following are the test results of the swear word detection system, presented in Table 1.

Table 1. The results of the swear word detection system

Which is expected / From System	Detected	Not detected
Detected	$TP = 11$	$FP = 3$
Not detected	$FN = 0$	$TN = 11$

After knowing the system test results, the research continued calculating the values of accuracy, precision, and recall with test results of 88%, 78.57%, and 100%. Because evaluation using precision and recall values is considered insufficient to indicate the quality of a system, the f measure is also used, which is the harmonic mean of precision and recall (Wahib et al., 2015). The results of the f measure measurement obtained a result of 88%, with the level of accuracy of the output generated by this swear word detection system not reaching 100%, because there are still detection errors, such as in the 6th test data, detected the word dog and garbage, the word should be The word was not detected as a swear word, because the use of the word in the 6th test data was as news explaining the anti-rabies vaccine for dogs and handling waste in watersheds. Furthermore, in the 8th test data, the word satan was detected, the word should not have been detected as a swear word because the word satan in the 8th test data came from the name of the Sesetan highway, which should be the

syllable of the word sesetan, not a prefix, but rather a prefix. one word. And in the 13th test data, the words pig and garbage were detected. These words should not have been detected as swear words because the use of these words in the 8th test data was as news explaining about fattening pigs and waste management involving poor people.

Discussion

The research results have found that the swear word detection system has detected various kinds of swear words on the Bali Province SKPD website. Swear word detection systems are widely used on various government websites. Swear words are harsh words and Sara (Sahrul & Rahman, 2019; Satiti & Ma'shumah, 2021). Words like this are certainly not worthy of seeing the general public. The website as a medium of information and publication for the public must, of course, present factual data in good and correct language (Afriansyah, 2020; Sugiartawan et al., 2018). Websites are generally used by various institutions, such as government and schools to provide information about institutions and services provided by these institutions (Mukti, 2018). Through a website page, an institution will be able to present information in the form of text, still or motion pictures, animation, sound, and or a combination of all of them, both static and dynamic, which form a series of interrelated buildings, each of which is connected by networks page (hyperlink) (Putra & Harli, 2021; Sugiartawan et al., 2018). A website designed by an institution will assist the process of delivering information to the public so that all administrative processes can be simplified.

Word detection on the website is carried out with several systems. One of which is the UML (Unified Modeling Language) system built using a web programming language with the Laravel framework. Unified Modeling Language is one of the visual modeling methods used in the design and manufacture of object-oriented software (Edde & Budayawan, 2021). UML contains writing standards or some blueprint that includes writing business process classes in a specific language (Marbun et al., 2021). UML as a language provides vocabulary and writing words in 'MS Word' for communication purposes (Prihandoyo, 2018). A model language is a language that has vocabulary and concepts of writing order and physically represents a system (Henderi, 2018).

Several studies that have been conducted previously revealed that the swear word detection system using the RNN model showed better performance than the ANN model with training, validation, and test accuracy of 94%, 84%, and 84%, respectively (Sahrul & Rahman, 2019). Another study also revealed that in addition to using the RNN model, swear words can also be detected using the SVM, Logistic Regression, and Naive Bayes models, which significantly succeeded in detecting insulting elements in comment texts from social media (Sastrodikoro et al., 2018). Further research revealed that the N-Gram method was able to see typing errors in non-standard words in Indonesian written works (Hartina & Masri, 2020). Based on some of the results of these studies, it can be said that detecting swear words has various methods and models. The word detection system can facilitate identifying swear words on the website.

4. CONCLUSION

Based on the results of the design and construction of a swear word detection system on the SKPD of Bali Province Government website content using the Nazief and Adriani stemming algorithm, as well as the Term Frequency weighting algorithm and also from the results of tests carried out using the accuracy measure test method, it can be concluded that the detection system testing swear words carried out using the accuracy measure test method, which uses test data with a total of twenty-five test data consisting of eleven news texts containing swear words and fourteen news texts that do not have to swear words, only

accuracy values are obtained by 88%; the precision value of 78.57%; recall value of 100%; and f measure of 88%, this is because the system has not been able to read the sentence structure, so that a word that should not be used as a swear word in a document/text becomes detected by the system as a swear word, for example, the word dog detected in the text of the news dog rabies.

5. REFERENCES

Afriansyah, R. (2020). Pembuatan Portal Website Sekolah Sma Negeri 1 Sungailiat Sebagai Media Informasi. *Dinamisia : Jurnal Pengabdian Kepada Masyarakat*, *5*(1). https://doi.org/10.31849/dinamisia.v5i1.4413.

Ari, D. P. S., & Hanum, L. (2021). Pengaruh Kualitas Pelayanan Website Djp Terhadap Kepuasan Pengguna Dengan Modifikasi E Govqual. *Jurnal Administrasi Bisnis*, *15*(01), 104–111. https://doi.org/10.21776/ub.profit.2021.015.01.11.

Ariana, putu bagus. (2020). Analisis Segitiga Kebijakan Kesehatan Perda Kawasan Tanpa Rokok Nomor 10 Tahun 2011 Provinsi Bali. *Jurnal Kesehatan STIKes Buleleng*, *4*(1), 74. https://doi.org/10.52073/midwinerslion.v4i1.127.

Edde, G. P., & Budayawan, K. (2021). Pembuatan Aplikasi Reminder Jadwal Perkuliahan di Jurusan Teknik Elektronika Berbasis Android. *Voteteknika (Vocational Teknik Elektronika Dan Informatika)*, *9*(4), 1. https://doi.org/10.24036/voteteknika.v9i4.112669.

Hartina, T., & Masri, A. (2020). Pendeteksi Kesalahan Pengetikan Kata Non Baku pada Karya Tulis Menggunakan Metode N-Gram. *Jurnal Informatika*, *7*(1), 77–84. https://doi.org/10.31311/ji.v7i1.7916.

Hayami, R., Fatma, Y., Soni, S., & Wenando, F. A. (2019). Pelatihan Pengelolaan Website Sebagai Upaya Meningkatkan Publikasi Profil Kelurahan Tangkerang Selatan Pekanbaru. *Jurnal Pengabdian UntukMu NegeRI*, *3*(2), 230–233. https://doi.org/10.37859/jpumri.v3i2.1619.

Henderi, D. (2018). Analisis Dan Perancangan Sistem Informasi Kepegawaian Menggunakan Unified Modeling Language (UML). *Jurnal Sistem Informasi Dan Teknologi Informasi*, *1*(7), 22–30. https://doi.org/10.36774/jusiti.v7i1.62.

Marbun, M., Situmorang, H., Hutagalung, D. M., & Sitanggang, R. (2021). Sistem Informasi Klinik Berbasis Website. *Jurnal Tekesnos*, *3*(1), 229–235. http://e-journal.sari-mutiara.ac.id/index.php/tekesnos/article/view/2281/1552.

Mubarak, A. (2019). Rancang Bangun Aplikasi Web Sekolah Menggunakan Uml (Unified Modeling Language) Dan Bahasa Pemrograman PHP (PHP Hypertext Preprocessor) Berorientasi Objek. *JIKO (Jurnal Informatika Dan Komputer)*, *2*(1), 19–25. https://doi.org/10.33387/jiko.v2i1.1052.

Mukti, Y. (2018). Rancang Bangun Website Sekolah Dengan Metode User Centered Design (UCD). *Jurnal Ilmiah Betrik*, *9*(02), 84–95. https://doi.org/10.36050/betrik.v9i02.34.

Nugraha, joko tri. (2018). E-Government Dan Pelayanan Publik (Studi Tentang Elemen Sukses Pengembangan E-Government Di Pemerintah Kabupaten Sleman). *Jurnal Komunikasi Dan Kajian Medai*, *2*(1). https://doi.org/10.31002/jkkm.v2i1.758.

Pradipa, N. A., & Munidewi, B. (2018). Pengaruh SPI Terhadap Kualitas Informasi Laporan Keuangan Dengan Komitmen Organisasi Dan Kompetensi SDM Sebagai Pemoderasi (Studi Pada SKPD Provinsi Bali). *Behavioral Accounting Journal*, *1*(1), 50–61. https://doi.org/10.33005/baj.v1i1.18.

Prihandoyo, M. T. (2018). Unified Modeling Language (UML) Model Untuk Pengembangan Sistem Informasi Akademik Berbasis Web. *Jurnal Informatika: Jurnal*

Pengembangan IT, *3*(1), 126–129. https://doi.org/ 10.30591/jpit.v3i1.765.

Putra, S. A., & Harli, E. (2021). Perancangan Sistem Informasi Koperasi Simpan Pinjam pada Sma Yaperjasa Berbasis Website. *Jurnal Riset Dan Aplikasi Mahasiswa Informatika (JRAMI)*, *2*(03). https://doi.org/10.30998/jrami.v2i03.1198.

Putri, K. M. R., Yuniarta, G. A., & Prayudi, M. A. (2017). Pengaruh Perencanaan Anggaran, Kualitas Sumber Daya Manusia Dan Komitmen Organisasi Terhadap Penyerapan Anggaran (Survei Pada SKPD di Wilayah Pemerintah Daerah Provinsi Bali). *Jurnal Ilmiah Akuntansi*, *8*(2). https://doi.org/ 10.23887/jimat.v8i2.13679.

Sahrul, S., & Rahman, A. (2019). Sistem Pendeteksi Kalimat Umpatan Di Media Sosial Dengan Model Neural Network. *Journal Untar*, *3*(2), 108–115. https://doi.org/10.24912/computatio.v3i2.6032.

Santoso, H. B., Delima, R., & Wibowo, A. (2019). Pelatihan Pengembangan Web Profil Desa bagi Aparatur Pemerintah Desa. *E-Dimas: Jurnal Pengabdian Kepada Masyarakat*, *10*(1), 41. https://doi.org/10.26877/e-dimas.v10i1.2592.

Sastrodikoro, Y. A., Palit, H. N., & Andjarwirawan, J. (2018). Aplikasi Pendeteksi Unsur Hinaan dalam Komentar di Media Sosial Berbahasa Indonesia. *Jurnal Infra Petra*, *6*(1). https://publication.petra.ac.id/index.php/teknik-informatika/article/view/6342.

Satiti, S. D., & Ma'shumah, N. K. (2021). Penggunaan Umpatan pada Siswa Sekolah Dasar di Kudus. *Sutasoma : Jurnal Sastra Jawa*, *9*(1), 68–83. https://doi.org/10.15294/sutasoma.v9i1.46162.

Setiawan, D. (2018). Dampak Perkembangan Teknologi Informasi dan Komunikasi Terhadap Budaya. *Jurnal Simbolika: Research and Learning in Communication Study*, *4*(1), 62. https://doi.org/10.31289/simbollika.v4i1.1474.

Sugiartawan, P., Rustina, I. D. K. R., & Saleh Insani, R. W. (2018). E-Government Media Informasi Alat Kelengkapan Dewan Provinsi Bali dan Media Diskusi Berbasis Website. *Jurnal Sistem Informasi Dan Komputer Terapan Indonesia (JSIKTI)*, *1*(2), 75–86. https://doi.org/10.33173/jsikti.17.

Tjahyanti, L. P. A. S. (2020). Pendeteksian Bahasa Kasar (Abusive Language) Dan Ujaran Kebencian (Hate Speech) Dari Komentar Di Jejaring Sosial. *Journal of Chemical Information and Modeling*, *07*(9), 1689–1699. https://ejournal.unipas.ac.id/index.php/DW/article/view/248/245.

Veza, O., & Nurlinda. (2021). Perancangan Media Pembelajaran Pengenalan Anggota Tubuh Manusia Dalam Bahasa Inggris Dan Bahasa Arab Berbasis Web Dan Multimedia Interaktif (Studi Kasus Taman Kanak-Kanak Al-Mi'raj Batam). *Jurnal Responsive Teknik Informatika*, *5*(01), 1–11. https://doi.org/10.36352/jr.v5i01.186.

Wahib, A., Pasnur, P., Santika, P. P., & Arifin, A. Z. (2015). Perangkingan Dokumen Berbahasa Arab Menggunakan Latent Semantic Indexing. *Jurnal Buana Informatika*. https://doi.org/10.24002/jbi.v6i2.411.

Widagdo, P. P., Haviluddin, H., Setyadi, H. J., Taruk, M., & Pakpahan, H. S. (2018). Sistem Informasi Website Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Mulawarman. *Prosiding SAKTI (Seminar Ilmu Komputer Dan Teknologi Informasi)*, *3*(2), 5–9. http://e-journals.unmul.ac.id/index.php/SAKTI/article/view/1818.

Yoga, S. (2019). Perubahan Sosial Budaya Masyarakat Indonesia Dan Perkembangan Teknologi Komunikasi. *Jurnal Al-Bayan*, *24*(1). https://doi.org/10.22373/albayan.v24i1.3175.