ISSN: 2454-1907
DOI: https://doi.org/10.29121/ijetmr.v6.i5.2019.380

IJETMR
International Journal of Engineering Technologies and Management Research
A Knowledge Repository

CLASSIFICATION OF NEWS TYPES BY IMPLEMENTING ENHANCED CONFIX STRIPPING STEMMER

Muhammad Ichwan Utari [1], Henny Medyawati [*2]
[1] Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia
[*2] Faculty of Economics, Gunadarma University, Indonesia

Abstract:
News has become a community need in the world. Managing a lot of news articles is not easy and takes a long time. Indonesia has various types of media platforms that display news, one of which is an online news portal. Automation systems that are capable of managing and grouping Indonesian language news articles are needed. This study designed and built a web-based application to classify types of Indonesian language news articles by implementing the Enhanced Confix Stripping Stemmer algorithm. The categories used in the system are entertainment, lifestyle, sports, technology, and economics. The data used is secondary data quoted from 2 online news portals in Indonesia. The system development method used is Rapid Application Development. The data used for testing amounts to 30 news. The average results obtained from the system accuracy test are 63%. This shows that the system performance for the classification of news types is good. The number of words in a news article is very influential during the classification process.

Keywords: News; Text Mining; Enhanced Confix Stripping Stemmer; Naïve Bayes Classifier.

Cite This Article: Muhammad Ichwan Utari, and Henny Medyawati. (2019). "CLASSIFICATION OF NEWS TYPES BY IMPLEMENTING ENHANCED CONFIX STRIPPING STEMMER." *International Journal of Engineering Technologies and Management Research*, 6(5), 135-141. DOI: https://doi.org/10.29121/ijetmr.v6.i5.2019.380.

1. Introduction

News has become a community need in the world. One of the media displays news, namely online news portals. News does not only refer to the press or the mass media, but on radio, television, film, and the internet. At first, the news only belonged to the newspaper. But now the news is attached to radio, television and the internet. Indonesia has many online news portals such as liputan6, tribunnews, kompas, detik, kompasiana, kapanlagi, and others.

The survey results of the Association of Indonesian Internet Service Providers (APJII) and Technologists stated that the growth of internet users in 2017 reached 143.26 million. Increased compared to 2016 amounting to 132.7 million people [1]. This will affect the growth and exchange of information. One of the effects is that news articles uploaded on the internet are very many with a fast span of time. Managing a lot of news articles is not easy and takes a long time. Automation systems that are capable of managing and grouping Indonesian language news articles are needed. This grouping is expected to simplify and streamline time in managing news documents.

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

Text mining is a way for text to be processed using a computer to produce useful analysis [2]. The stages of text mining are generally divided into several general stages, namely, preprocessing, feature selection, and stemming [3]. Stemming is the process of mapping and decomposing various forms of a word into their basic words. The purpose of stemming is to eliminate affixes in the form of prefixes, suffixes, and confixes to each word. Indonesian has morphological rules, so the stemming process must be based on the Indonesian morphological rules [3].

Stemming has several methods, one of which will be used in this study is Enhanced Confix Stripping Stemmer (ECS). ECS is a stemming method in Indonesian which was introduced by AgusZainalArifin, I PutuKertaMahendra, Henning TitiCiptaningtyas in 2014. ECS is a development of the Confix Stripping Stemmer (CS) method introduced by Jelita Asian in 2007 [4].

2. Materials and Methods

2.1. Enhanced Confix Stripping (ECS) Stemmer

Enhanced Confix Stripping Stemmer is the result of the development of Stemmer Confix Stripping conducted by AgusZainalArifin, I PutuKertaMahendra, HenningTitiCiptaningtyas in 2014. Based on the failures of Confix Stripping Stemmer, Arifin, Mahendra, and Ciptaningtyas trying to increase Confix Stripping Stemmer, and presents a modified Confix Stripping Stemmer called Enhanced Confix Stripping Stemmer. The repair agreement is as follows:

1) Modifying some of the rules in Tables 2 and 3, so that the process of stemming the words with construction "mem + p ...", "male + s ...", "meng + +", "peng + +. ... ", and" peng + k ... "can be done. These modifications are listed in Table 6.
2) Add additional stemming steps to resolve the problem of deleting suffixes. This additional step is called LoopPengembalianAkhiran. This step is done when recording (step Confix Stripping Stemmer) fails [4].

At each LoopPengembalianAkhiran process, a dictionary search is performed to check the results in the current word. The process in the loopPengembalianAkhiran is defined as follows:

1) Revert the word to the pre-encoding form and return all prefixes that were deleted in the last process, so that it will create the word model like this:

   [DP + [DP + [DP]]] + Term

   The deletion of the prefix is tried. If the dictionary search is successful, the process stops. If not, the next step is executed.
2) Return the ending that was deleted earlier. This means that the return starts from DS ("-i", "-kan", "-an") if it exists, then is followed by PP ("- ku", "-mu", "-nya"), and finally is P ("-lah", "-kah", "-ah", "-pun"). On each return, steps 3 to 5 are tried. A special case for DS "-kan", the character "k" is restored first and steps 3 to 5 are executed. If it still fails, then "an" is restored.
3) Prefixes removal is carried out according to the rules defined in Tables 2, 3, 4, and 5 (with modifications in Table 6).

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

4) Recoding.
5) If the dictionary search is not successful, return the word to the pre-encoding form and return all the prefixes that were deleted. The next suffix in the order in step 1 is restored and steps 3 to 5 are performed on the current word [4].

Table 1: The flow of prefixes removal for the "me-" prefix

Rule	Construction	Prefix Removal
1	me{l|r|w|y}V...	me-{l|r|w|y}V...
2	mem{b|f|v}...	mem-{b|f|v}...
3	mempe...	mem-pe...
4	mem{rV|V}...	me-m{rV|V}... | me-p{rV|V}...
5	men{c|d|j|z}...	men-{c|d|j|z}...
6	menV...	me-nV... | me-tV
7	meng{g|h|q|k}...	meng-{g|h|q|k}...
8	mengV...	meng-V... | meng-kV...
9	menyV...	meny-sV...
10	mempV...	mem-pV... where V!= 'e'

Table 2: The flow of prefixes removal for the "pe-" prefix

Rule	Construction	Prefix Removal
1	pe{w|y}V...	pe-{w|y}V...
2	perV...	per-V... | pe-rV...
3	perCAP	per-CAP... where C!= 'r' and P!= "er"
4	perCAerV...	per-CAerV... where C!= 'r'
5	pem{b|f|V}...	pem-{b|f|V}...
6	pem{rV|V}...	pe-m{rV|V}... | pe-p{rV|V}...
7	pen{c|d|j|z}...	pen-{c|d|j|z}...
8	penV...	pe-nV... | pe-tV...
9	peng{g|h|q}...	peng-{g|h|q}...
10	pengV...	peng-V... | peng-kV...
11	penyV...	peny-sV...
12	pelV...	pe-lV... except "pelajar", return "ajar"
13	peCerV...	per-erV... where C!={r|w|y|l|m|n}
14	peCP...	pe-CP... where C!={r|w|y|l|m|n} and P!= 'er'
15	terC$_1$erC$_2$...	ter-C$_1$erC$_2$... where C1!='r'
16	peC$_1$erC$_2$...	pe-C$_1$erC$_2$... where C$_1$!={r|w|y|l|m|n}

Table 3: The flow of prefixes removal for the "be-" prefix

Rule	Construction	Prefix Removal
1	berV...	ber-V... | be-rV...
2	berCAP...	ber-CAP... where C!= 'r' and P!= "er"
3	berCAerV...	ber-CaerV... where C!= 'r'
4	belajar	bel-ajar
5	beC$_1$erC$_2$...	be-C$_1$erC$_2$... where C$_1$!={ 'r'|'l'}

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

Table 4: The flow of prefixes removal for the "te-" prefix

Rule	Construction	Prefix Removal
1	terV...	ter-V... | te-rV...
2	terCerV...	ter-CerV... where C!= 'r'
3	terCP...	ter-CP... where C!= 'r' and P!= 'er'
4	teC$_1$erC$_2$...	te-C$_1$erC$_2$... where C$_1$!= 'r'

Table 5: Flow modification for Table 2

Rule	Construction	Prefix Removal
1	men{c|d|j|s|z}...	men-{c|d|j|s|z}...
2	mengV...	meng-V... | meng-kV... | (mengV-... if V='e')
3	mempA...	mem-pA... where A!= 'e'

Table 6: Flow modification for Table 3

Rule	Construction	Prefix Removal
1	pengC...	peng-C...
2	pengV...	peng-V... | peng-kV... | (pengV-... if V='e')

2.2. Rapid Application Development(RAD)

The method used to design and build a system to implement Enhanced Confix Stripping Stemmer (ECS) in the classification of news types is Rapid Application Development (RAD). Rapid Application Development (RAD) is a system development strategy that emphasizes the speed in development through user involvement in development quickly, iteratively, and incrementally from a series of prototypes and a system that can develop into a final system or a particular version [5].

The RAD stages are as follows:

- Requirement planning.
- RAD Design Workshop.
- Construction.
- Implementation.

The type of test data used is secondary data in the form of news articles sourced from the news portal kompas.com and cnnindonesia.com. The number of news articles used is 30 news articles. Data is taken by quoting a news article that has been published on a news portal that has been determined. The news categories used are economics, technology, entertainment, lifestyle, and sports.

3. Results and Discussions

Based on some previous studies Enhanced Confix Stripping Stemmer algorithm is an algorithm that has the highest level of accuracy compared to other stemming algorithms such as Porter Stemmer, Nazief-Adriani, and Confix Stripping Stemmer. The system to be built is implementing the Enhanced Confix Stripping Stemmer Algorithm in the form of a web-based application to classify the types of news articles. This system is expected to provide high accuracy in terms of

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

the classification of news types. Users only need to upload news on the system, then the news will be automatically classified by this system.

The scope of the system built is as follows:

- The categories of news used are five categories, namely entertainment, lifestyle, sports, technology, and economy. The news data is obtained from online news portals.
- The news used in this study is Indonesian language news.
- In the text mining stage, the tagging stage is not carried out because it does not handle English-language texts.
- The system built is not integrated with existing online news portals, but by making its own homepage web-based and using offline networks.

Testing the results of the classification is done to determine the level of accuracy of the system implementation of the Enhanced Confix Stripping Stemmer algorithm for classification of news types. Tests are carried out on class results for test data. The results of the testing carried out by the system are shown in Table 7. The testing model is comparing the test news categories available on online news portals that have been determined with the system that has been built.After calculating the accuracy of 5 iterations, then calculate the average accuracy. Calculation of the accuracy of the implementation of Enhanced Confix Stripping Stemmer algorithms is shown in Table 8.

Table 7: Classification Testing Table

No	News Title	News Portal	Testing	Results	Minutes
1	JK Sebut Bank TakAdilBeriBungaTinggikePengusaha Kecil	Economy	Economy	Succeeded	4,7
2	Darmin Ragu DP NolPersenDongkrakPembiayaanMultifinance	Economy	Economy	Succeeded	7
3	BagasiBerbayarDisebutBakalKerekInflasi	Economy	Economy	Succeeded	3,2
4	HargaKaretTakWajar, DarminAjakBicara Bursa Internasional	Economy	Economy	Succeeded	4,4
5	HargaPengirimanPaketLewat JNE NaikMulai 15 Januari 2019	Economy	Technology	Failed	2,5
6	Hero Supermarket KlaimPenuhiHak PHK 92 PersenKaryawan	Economy	Economy	Succeeded	3,9
7	Amazon InvestasiKembangkanTrukAngkutOtonom	Technology	Economy	Failed	2,6
8	APM PelajariAturanBaruBeli Mobil DP 0 Persen	Technology	Economy	Failed	4,2
9	Google DiizinkanUniEropaBatasiHakuntukDilupakan	Technology	Entertainment	Failed	5
10	Apple BakalLuncurkanFiturKamera di iPhone AnyarTahunIni	Technology	Economy	Failed	3,3
11	KemenhubKlarifikasi, Tidak Ada Uji KIR BuatOjek Online	Technology	Technology	Succeeded	3,8
12	CEO NvidiaSebutHukum Moore TelahMati	Technology	Economy	Failed	8,3
13	4 PesepakbolaBercodetKhas, TermasukLescott The klingon	Sports	Lifestyle	Failed	3,6
14	FitrianiJuara Thailand Masters 2019	Sports	Sports	Succeeded	4,4

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

15	KhabibBalasTuduhan McGregor di MedsosTerkait Duel UFC 229	Sports	Entertainment	Failed	3,6
16	AsistenPelatihPersijaPuji Bruno Matos di LagaAmal Lampung	Sports	Sports	Succeeded	4,5
17	Ducati LebihHarmonisTanpa Jorge Lorenzo	Sports	Sports	Succeeded	3,6
18	Bantai Benevento, Inter Lolos kePerempat Final Coppa Italia	Sports	Sports	Succeeded	5,1
19	Dhira Bongs SampingkanBekrafuntukTampil di SXSW 2019	Entertainment	Entertainment	Succeeded	8,3
20	DuaLipaDihadirkandalamWujudPatungLilin	Entertainment	Entertainment	Succeeded	3,5
21	Jake GyllenhaalHadapiLukisanBerhantu di Velvet Buzzsaw	Entertainment	Entertainment	Succeeded	3,6
22	Al Pacino DisebutBergabung di Drama TV The Hunt	Entertainment	Entertainment	Succeeded	4,6
23	Ceraidari Bezos, MacKenzieBakalJadiWanitaTerkayaDunia	Entertainment	Lifestyle	Failed	5,3
24	Final Destination 6 DigarapOlehPenulisSaw	Entertainment	Entertainment	Succeeded	3,2
25	Cara MencegahdanMenghilangkanLemak di Perut	Lifestyle	Lifestyle	Succeeded	4,6
26	Agenda AkhirPekanMingguIni	Lifestyle	Entertainment	Failed	5,7
27	Cara CantikPilihLipstik	Lifestyle	Lifestyle	Succeeded	5,5
28	4 TrikCerdasBeliOleh-olehSaatLiburan	Lifestyle	Lifestyle	Succeeded	4,2
29	6 Hal yang WajibDiperhatikanSaatPilih Sepatu Boots	Lifestyle	Lifestyle	Succeeded	6,7
30	KemenkesSebut Tender Obat HIV DimulaiBulanDepan	Lifestyle	Economy	Failed	6,1

Table 8: Accuracy Calculation Results

Fold	**Accuracy**
I	83%
II	17%
III	67%
IV	83%
V	67%
Average accuracy	63%

Based on Table 7, the average accuracy is 63% as the final accuracy of the implementation of the Enhanced Confix Stripping Stemmer algorithm. This means that the performance of the Enhanced Confix Stripping Stemmer algorithm applied in the problem of text mining implementation for news type classification is good enough.

The dictionary of basic words in the system database greatly influences the performance of the Enhanced Confix Stripping Stemmer algorithm. The time of the classification process is influenced by the number of words in the test news. The average time obtained from the testing process is 4.5 minutes.

ISSN: 2454-1907
DOI: 10.5281/zenodo.3232918

4. Conclusions & Recommendations

4.1. Conclusions

Based on the results and discussion, the conclusion that can be drawn is that the website classifies the type of news content by implementing Enhanced Confix Stripping Stemmer (ECS) as a word stemming algorithm that has been successfully built.

4.2. Recommendations

The following are some suggestions for further development of this research:

- It is expected that in further research, this system can be developed again using English text as test data.
- Further research is expected to be able to use programming languages besides PHP languages, such as Java, C ++, and others.
- Further research can try to compare the performance of text classification using Enhanced Confix Stripping Stemmer Algorithm and Naïve Bayes Classifier, with other stemming and classifier algorithms.

References

[1] APJII (2017). HasilSurveiPenetrasidanPerilakuPengguna Internet Indonesia 2017. AsosiasiPenyelenggaraJasa Internet Indonesia. [Blog online]. NN. Available from: https://apjii.or.id/content/read/39/342/Hasil-Survei-Penetrasi-dan-Perilaku-Pengguna-Internet-Indonesia-2017[Accessed 18 Mei 2018].

[2] Witten, I.A., Frank, E., Hall, M.A. (2011). Data Mining Practical Machine Learning Tools and Techniques. USA: Elsevier.

[3] Triawati, C. (2009). MetodePembobotan Statistical Concept Based untukKlasteringdanKategorisasiDokumenBerbahasa Indonesia. Bandung: InstitutTeknologi Telkom.

[4] Arifin, A.Z., Mahendra, I.P.A.K., Ciptaningtyas, H.T. (2014). Enhanced Confix Stripping Stemmer And Ants Algorithm For Classifying News Document In Indonesian Language. In: The 5th International Conference on Information & Communication Technology and Systems, 2014. Irbid. Jordan. pp. 149-158.

[5] Bentley, L.D., Whitten, J.L. (2007). System Analysis and Design Methods. New York: McGraw-Hill/Irwin.

*Corresponding author.
E-mail address: henmedya@ staff.gunadarma.ac.id