**README FIRST**

This corpus is derived from several freely accessible Indonesian news website for research purpose only. The news websites are:

* kompas.com is a registered trademark of PT. Kompas Cyber Media. <https://inside.kompas.com/about-us>
* tempo.co is a registered trademark of PT INFO MEDIA DIGITAL. <https://www.tempo.co/about>
* merdeka.com is a registered trademark of PT KAPAN LAGI DOT COM NETWORKS. <https://www.merdeka.com/company/tentang-kami.html>
* republika.co.id is a registered trademark of PT Republika Media Mandiri. <https://www.republika.co.id/page/about>
* viva.co.id is a registered trademark of PT. Viva Media Baru. <https://www.viva.co.id/tentang-kami>
* tribunnews.com is a registered trademark of PT Tribun Digital Online. http://www.tribunnews.com/about-us

The corpus is a part of bachelor thesis work of Aad Miqdad Muadz Muzad. Please cite this publication if you use this corpus in your research:

MUZAD, Aad Miqdad Muadz; RAHUTOMO, Faisal. Korpus Berita Daring Bahasa Indonesia Dengan Depth First Focused Crawling. **Prosiding Sentrinov (Seminar Nasional Terapan Riset Inovatif)**, [S.l.], v. 2, n. 1, p. 11-20, oct. 2016. ISSN 2477-2097. Available at: <<http://proceeding.sentrinov.org/index.php/sentrinov/article/view/163>>.

We crawled several categories of the websites for 6 months from July 2015 until December 2015. The corpus contains 150,466 news articles. The authors are not responsible for any further use of this corpus against laws.

The corpus is stored in JSON and XML format. The JSON format has this structure:

{

“sumber”: “”,

“Tanggal”: “”,

“kategori”: “”,

“judul”: “”,

“isi”: “”,

“link”: “”

}

Table 1. Corpus information in detail

| **Month** | **Website** | **Category** | **Articles** | **Words** | **Average words/ article** |
| --- | --- | --- | --- | --- | --- |
| Juli 2015 | kompas.com | Teknologi | 328 | 95132 | 290,0365854 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 604 | 151755 | 251,25 |
| Nasional | 1787 | 491215 | 274,8824846 |
| Olahraga | 236 | 56621 | 239,9194915 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 335 | 101072 | 301,7074627 |
| republika.co.id | Teknologi | 190 | 36460 | 191,8947368 |
| Otomotif | 37 | 8276 | 223,6756757 |
| Bisnis Ekonomi | 1202 | 245014 | 203,8386023 |
| Nasional | 5576 | 844197 | 151,3983142 |
| Olahraga | 133 | 17532 | 131,8195489 |
| Bola | 1845 | 196089 | 106,2813008 |
| Lifestyle | 811 | 160013 | 197,3033292 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 317 | 89379 | 281,9526814 |
| Otomotif | 463 | 99505 | 214,9136069 |
| Bisnis Ekonomi | 904 | 213521 | 236,1957965 |
| Nasional | 1985 | 469721 | 236,6352645 |
| Olahraga | 279 | 63027 | 225,9032258 |
| Bola | 1875 | 367724 | 196,1194667 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 128 | 27492 | 214,78125 |
| Bisnis Ekonomi | 707 | 142433 | 201,4611033 |
| Nasional | 2647 | 506624 | 191,3955421 |
| Olahraga | 497 | 94933 | 191,0120724 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 253 | 55679 | 220,0750988 |
| Travel | 609 | 83144 | 136,5254516 |
| Agustus 2015 | kompas.com | Teknologi | 344 | 95042 | 276,2848837 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 883 | 232924 | 263,7870895 |
| Nasional | 1811 | 499964 | 276,0706792 |
| Olahraga | 286 | 78244 | 273,5804196 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 371 | 114641 | 309,0053908 |
| republika.co.id | Teknologi | 289 | 52792 | 182,6712803 |
| Otomotif | 98 | 18279 | 186,5204082 |
| Bisnis Ekonomi | 1340 | 222342 | 165,9268657 |
| Nasional | 4103 | 596105 | 145,2851572 |
| Olahraga | 221 | 27022 | 122,2714932 |
| Bola | 1962 | 227834 | 116,1233435 |
| Lifestyle | 1032 | 131661 | 127,5784884 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 405 | 114143 | 281,8345679 |
| Otomotif | 648 | 148398 | 229,0092593 |
| Bisnis Ekonomi | 1165 | 275575 | 236,5450644 |
| Nasional | 1810 | 455581 | 251,7022099 |
| Olahraga | 377 | 87888 | 233,1246684 |
| Bola | 2170 | 437461 | 201,5949309 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 315 | 75372 | 239,2761905 |
| Bisnis Ekonomi | 1014 | 208820 | 205,9368836 |
| Nasional | 2981 | 596491 | 200,0976182 |
| Olahraga | 626 | 117808 | 188,1916933 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 240 | 51153 | 213,1375 |
| Travel | 579 | 76649 | 132,3816926 |
| September 2015 | kompas.com | Teknologi | 396 | 115761 | 292,3257576 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 942 | 246274 | 261,4373673 |
| Nasional | 2029 | 558441 | 275,2296698 |
| Olahraga | 195 | 52779 | 270,6615385 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 412 | 147738 | 358,5873786 |
| republika.co.id | Teknologi | 268 | 51457 | 192,0037313 |
| Otomotif | 69 | 13260 | 192,173913 |
| Bisnis Ekonomi | 1185 | 203243 | 171,5130802 |
| Nasional | 4398 | 627014 | 142,5679854 |
| Olahraga | 140 | 17845 | 127,4642857 |
| Bola | 1985 | 211069 | 106,3319899 |
| Lifestyle | 914 | 143436 | 156,9321663 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 387 | 114991 | 297,1343669 |
| Otomotif | 557 | 124542 | 223,5942549 |
| Bisnis Ekonomi | 1143 | 270145 | 236,3473316 |
| Nasional | 2005 | 514408 | 256,5625935 |
| Olahraga | 370 | 85182 | 230,2216216 |
| Bola | 2266 | 471598 | 208,1191527 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 117 | 24481 | 209,2393162 |
| Bisnis Ekonomi | 1092 | 212993 | 195,0485348 |
| Nasional | 3003 | 579482 | 192,967699 |
| Olahraga | 530 | 104677 | 197,5037736 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 302 | 57251 | 189,5728477 |
| Travel | 492 | 69233 | 140,7174797 |
| Oktober 2015 | kompas.com | Teknologi | 386 | 104662 | 271,1450777 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 786 | 209258 | 266,2315522 |
| Nasional | 1938 | 521171 | 268,9220846 |
| Olahraga | 270 | 79748 | 295,362963 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 402 | 136420 | 339,3532338 |
| republika.co.id | Teknologi | 251 | 50620 | 201,6733068 |
| Otomotif | 105 | 16532 | 157,447619 |
| Bisnis Ekonomi | 1256 | 174085 | 138,602707 |
| Nasional | 5391 | 741706 | 137,5822667 |
| Olahraga | 218 | 22131 | 101,5183486 |
| Bola | 1903 | 184265 | 96,82869154 |
| Lifestyle | 974 | 180659 | 185,4815195 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 398 | 114463 | 287,5954774 |
| Otomotif | 538 | 119800 | 222,6765799 |
| Bisnis Ekonomi | 1110 | 270821 | 243,9828829 |
| Nasional | 1993 | 512748 | 257,2744606 |
| Olahraga | 374 | 87150 | 233,0213904 |
| Bola | 2065 | 430469 | 208,4595642 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 135 | 26147 | 193,6814815 |
| Bisnis Ekonomi | 1002 | 182697 | 182,3323353 |
| Nasional | 3056 | 547963 | 179,3072644 |
| Olahraga | 624 | 117661 | 188,5592949 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 257 | 44253 | 172,1906615 |
| Travel | 434 | 60388 | 139,1428571 |
| November 2015 | kompas.com | Teknologi | 381 | 101254 | 265,7585302 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 908 | 231669 | 255,1420705 |
| Nasional | 1727 | 456738 | 264,4690214 |
| Olahraga | 198 | 53137 | 268,3686869 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 316 | 102733 | 325,1044304 |
| republika.co.id | Teknologi | 353 | 64549 | 182,8583569 |
| Otomotif | 93 | 15515 | 166,827957 |
| Bisnis Ekonomi | 982 | 133949 | 136,404277 |
| Nasional | 5231 | 648433 | 123,9596635 |
| Olahraga | 282 | 33010 | 117,0567376 |
| Bola | 1694 | 187595 | 110,7408501 |
| Lifestyle | 842 | 131002 | 155,584323 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 400 | 117699 | 294,2475 |
| Otomotif | 516 | 117670 | 228,0426357 |
| Bisnis Ekonomi | 1088 | 270279 | 248,4181985 |
| Nasional | 1555 | 396433 | 254,940836 |
| Olahraga | 323 | 75796 | 234,6625387 |
| Bola | 1849 | 394983 | 213,6197945 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 91 | 18113 | 199,043956 |
| Bisnis Ekonomi | 1045 | 190975 | 182,7511962 |
| Nasional | 2994 | 534933 | 178,6683367 |
| Olahraga | 646 | 118555 | 183,5216718 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 187 | 36527 | 195,3315508 |
| Travel | 491 | 68670 | 139,8574338 |
| Desember 2015 | kompas.com | Teknologi | 288 | 75076 | 260,6805556 |
| Otomotif | 0 | 0 | 0 |
| Bisnis Ekonomi | 925 | 230374 | 249,052973 |
| Nasional | 1870 | 447518 | 239,3144385 |
| Olahraga | 159 | 40826 | 256,7672956 |
| Bola | 0 | 0 | 0 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 296 | 104674 | 353,6283784 |
| republika.co.id | Teknologi | 444 | 66963 | 150,8175676 |
| Otomotif | 97 | 12581 | 129,7010309 |
| Bisnis Ekonomi | 943 | 147443 | 156,3552492 |
| Nasional | 6150 | 796517 | 129,5149593 |
| Olahraga | 207 | 23466 | 113,3623188 |
| Bola | 1674 | 165417 | 98,81541219 |
| Lifestyle | 790 | 126905 | 160,6392405 |
| Travel | 0 | 0 | 0 |
| viva.co.id | Teknologi | 351 | 104724 | 298,3589744 |
| Otomotif | 555 | 121881 | 219,6054054 |
| Bisnis Ekonomi | 975 | 242893 | 249,1210256 |
| Nasional | 1622 | 389123 | 239,9032059 |
| Olahraga | 292 | 65048 | 222,7671233 |
| Bola | 1978 | 411687 | 208,1329626 |
| Lifestyle | 0 | 0 | 0 |
| Travel | 0 | 0 | 0 |
| tribunnews.com | Teknologi | 0 | 0 | 0 |
| Otomotif | 94 | 18609 | 197,9680851 |
| Bisnis Ekonomi | 845 | 161148 | 190,7076923 |
| Nasional | 3166 | 543016 | 171,5148452 |
| Olahraga | 394 | 75327 | 191,1852792 |
| Bola | 1592 | 246606 | 154,9032663 |
| Lifestyle | 168 | 28627 | 170,3988095 |
| Travel | 518 | 78093 | 150,7586873 |
| TOTAL | | | 150466 | 28736623 | 190,9841625 |