Indonesian Headline Detection on Twitter Social Media

Kaenova Mahendra Auditama
School of Computing
Telkom Univeristy
Bandung, Indonesia
kaenova@student.telkomuniversity.ac.id

Mahendra Dwifebri Purbolaksono
School of Computing
Telkom University
Bandung, Indonesia
mahendradp@telkomuniversity.ac.id

Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

***Abstrak*— Artikel media sosial seperti tweet di Twitter adalah salah satu sumber utama analisis teks. Namun, sejumlah besar data yang dikumpulkan seringkali mengandung kesalahan, seperti artikel yang tidak relevan yang mempengaruhi ketidakakuratan sistem. Salah satu yang dapat mempengaruhi ketidakakuratan sistem adalah banyaknya teks headline. Teks headline adalah bagian dari teks yang menjelaskan teks di bawahnya. Oleh karena itu, penelitian ini merancang metode yang efektif untuk mengumpulkan teks headline dan teks bukan headline, serta membangun model untuk mendeteksi teks headline. Metode pengumpulan data yang dikembangkan memiliki tingkat kesalahan yang rendah sehingga menjadi acuan yang baik untuk mendapatkan teks headline dan teks bukan headline. Model yang dibangun untuk mendeteksi teks headline memiliki kinerja yang baik dengan akurasi tertinggi yang dicapai oleh model IndoBERTweet sebesar 0,9921. Model CNN dan fastText yang dibangun memiliki akurasi yang lebih rendah masing-masing 0,9081 dan 0,8793, tetapi 200-2000x lebih cepat daripada IndoBERTweet. Selain itu, temuan kami mengungkapkan bahwa lebih dari setengah total data dalam empat topik terkait COVID-19 terdiri dari teks headline. Selanjutnya, kesalahan yang diamati dalam model dapat dikaitkan dengan sifat intrinsik yang kompleks dari teks headline, keterbatasan fastText dalam merepresentasikan kata-kata yang tidak standar, dan data pelatihan yang digunakan untuk membangun model fastText.**

Keywords—deteksi, headline, Twitter, IndoBERTweet, CNN, fastText