# BAB I
# PENDAHULUAN

## 1.1. Latar Belakang

Di era serba digital saat ini, semua orang pasti sudah tidak asing lagi dengan teknologi. Perkembangan teknologi yang begitu pesat sangat membantu masyarakat dalam melakukan segala aktivitasnya terkhusus dalam mengakses informasi. Setiap orang pastinya membutuhkan informasi untuk membantu mereka dalam mempertimbangkan pengambilan keputusan. Berdasarkan laporan digital Global (DataReportal, 2022) dari total penduduk dunia sebesar 7,91 miliar, terdapat 4,95 miliar pengguna internet dan 4,62 miliar diantaranya adalah pengguna media sosial. Selain itu, disebutkan juga bahwa pengguna media sosial rata -rata menghabiskan waktu selama 2 jam 27 menit setiap harinya hingga menciptakan volume data yang sangat besar. Media sosial menjadi salah satu wadah yang cukup diminati oleh generasi millennial dalam mengakses berbagai macam informasi. Karena penggunaan media sosial itu sendiri dianggap lebih mudah dan praktis untuk memenuhi setiap informasi yang mereka butuhkan (Meilinda, 2018; Putri & Irhandayaningsih, 2021). Selain itu, berdasarkan hasil survey dari *Katadata Insight Center* (KIC), 73 persen pengguna Internet di Indonesia cenderung memanfaatkan media sosial sebagai sumber informasi yang paling mudah diakses (Vania, 2022). Namun, seberapa banyak atau sejauh mana fakta yang beredar diverifikasi masih menjadi pertanyaan besar. Pada kenyataannya kemudahan untuk mengakses dan menyebarluaskan informasi seringkali juga dimanfaatkan oleh oknum tidak bertanggungjawab untuk menyebarkan hoaks.

Menurut Ketua Komunitas Masyarakat Anti Fitnah Indonesia (MAFINDO), Septiaji Eko Nugroho menjelaskan bahwa hoaks merupakan informasi yang direkayasa untuk menutup-nutupi kebenaran informasi yang sebenarnya. Selain itu, hoaks juga merupakan upaya untuk memutarbalikkan fakta dengan menggantinya dengan informasi-informasi yang meyakinkan tetapi tidak dapat diverifikasi kebenarannya (Kurniasih, 2021). Hoaks umumnya berisi teori konspirasi yang tidak dapat terbukti kebenarannya serta memuat penyataan palsu yang bias atau menyesatkan dengan maksud untuk menggiring opini pembaca kepada asumsi

negatif ataupun membodohkan publik. Berita hoaks tidak selalu dimulai dengan niat yang salah, beberapa mungkin berasal dari pendapat yang disalahartikan atau kebenaran yang disalahpahami (Brisola & Doyle, 2019). Meskipun demikian, berita hoaks tetap berdampak membahayakan dan tidak dapat diabaikan. Fenomena penyebaran berita hoaks menjadi populer di era digital saat ini, dengan bantuan Internet dan kemudahan yang diberikan oleh media sosial dalam mengunggah dan membagikan informasi, memudahkan peredaran dan replikasi berita hoaks menjadi sangat cepat viral. Banyaknya informasi yang bermunculan di media sosial membuat pengguna kewalahan oleh cepatnya jumlah informasi yang terus meningkat sehingga pengguna tidak punya waktu untuk memeriksa asal, kredibilitas, dan kebenaran informasi yang mereka terima.

Menurut beberapa studi menilai bahwa pendeteksian berita hoaks secara otomatis menggunakan teknik *deep learning* dapat mengatasi keterbatasan masyarakat dalam mengenali berita hoaks secara cepat dan akurat, sehingga dapat menjadi solusi pencegahan penyebaran berita hoaks. *Deep learning* adalah konsep *machine learning* berdasarkan jaringan saraf buatan (Janiesch et al., 2021). *Convolutional Neural Network* (CNN) dan *Recurrent Neural Network* (RNN), dua jenis utama arsitektur *Deep Neural Network* (DNN) atau jaringan saraf mendalam, banyak dieksplorasi untuk menangani berbagai tugas *Natural Language Processing* (NLP) (Yin et al., 2017), termasuk dalam tugas klasifikasi teks. Meskipun CNN sering dimanfaatkan untuk menyelesaikan permasalahan *computer vision*, akan tetapi dalam beberapa kasus tertentu CNN dapat menandingi kinerja model RNN. CNN seharusnya lebih baik dalam mengekstraksi posisi berbagai fitur terkait sementara RNN memodelkan secara berurutan ketergantungan jangka panjang (Zulqarnain et al., 2020). Tapi, dikarenakan arsitektur RNN sederhana seringkali mengalami permasalahan *vanishing gradient*, maka arsitektur seperti *Long Short-Term Memory* (LSTM) dari varian RNN diajukan untuk menangani masalah ini.

Korpora merupakan sumber daya, bank, atau sekumpulan teks dalam satuan apapun (kata, frasa, klausa, dan lain-lain) baik dalam bentuk tulisan maupun lisan. Semakin besar ukuran korpora, maka akan semakin banyak pula kosakata yang dimiliki sehingga pembelajaran pada model menjadi lebih efektif dalam mengenali keanekaragaman bahasa. Dalam perkembangan penelitian NLP, bahasa Indonesia

dianggap kurang terwakilkan dan termasuk bahasa yang dikategorikan sebagai *low resource language*, yang berarti bahasa Indonesia memiliki sumber daya atau kumpulan data yang rendah. Sedikitnya ketersediaan jumlah korpora yang besar dalam bahasa Indonesia dan kurangnya dokumentasi untuk melatih model *machine learning* secara efektif berpotensi menghambat penelitian yang akan dilakukan dan dapat mempengaruhi kinerja model nantinya. Sedangkan untuk membangun sebuah kumpulan data dibutuhkan sumber daya dan usaha yang besar, serta arsitektur komputer dengan kinerja yang tinggi (Suadaa et al., 2021). Sementara itu, berkaitan dengan proses pelatihan model CNN maupun LSTM yang digunakan dalam penelitian, keduanya membutuhkan kumpulan data yang cukup besar agar model yang dilatih dapat memiliki kinerja yang baik. Oleh sebab itu, penggunaan *pretrained language models* melalui metode *transfer learning* diajukan untuk mengatasi permasalahan tersebut. Model yang telah dilatih dapat menyimpan pengetahuan dari pembelajaran sebelumnya, kemudian dengan *transfer learning* pengetahuan tersebut dipindahkan ke model lainnya untuk menyelesaikan tugas NLP yang serupa. *Transfer learning* dapat menjadi titik awal yang bagus dalam peningkatan pembelajaran pada model apabila tidak memiliki jumlah sumber daya yang besar melalui transfer pengetahuan dari tugas yang sudah dipelajari sebelumnya.

Terkait hal tersebut, belum lama ini terdapat dua versi *monolingual pretrained language model* dengan arsitektur BERT yang diluncurkan bersamaan pada tahun 2020 bernama IndoBERT. Versi pertama IndoBERT pertama kali diajukan oleh tim IndoNLU, yang telah dilatih pada dataset Indo4B yang dikumpulkan dari sumber yang tersedia untuk umum seperti teks media sosial, blog, berita dan situs (Wilie et al., 2020). Lalu versi kedua dari IndoBERT diajukan oleh tim yang diprakarsai Fajri Koto, yang telah dilatih pada dataset IndoLEM yang dikumpulkan dari Wikipedia Indonesia, artikel berita (Kompas, Tempo, and Liputan6), dan situs korpora (Koto et al., 2020). Kedua monolingual BERT ini dilatih pada korpora berbahasa Indonesia untuk mendorong penelitian lebih lanjut dalam *transfer learning* ke berbagai tugas pemrosesan bahasa Indonesia (Suadaa et al., 2021).

Kemudian, melihat potensi besar dari teknik *transfer learning* yang sudah disebutkan, pada kesempatan kali ini peneliti ingin menguji sejauh mana keefektifan dari teknik *transfer learning* dengan mengkombinasikan *pretrained language model* IndoBERT pada model CNN maupun LSTM sebagai lapisan yang mampu memberikan representasi yang berbeda dibandingkan dengan menggunakan lapisan *embedding* pada umumnya.

Berdasarkan uraian latar belakang di atas, meskipun penerapan teknik *deep learning* seperti CNN dan LSTM mampu mengatasi permasalahan klasifikasi berita hoaks secara efektif, namun keterbatasan data yang dimiliki menjadi hambatan dalam menghasilkan kinerja yang baik. Sehingga perlu dilakukan penerapan teknik *transfer learning* dan penelitian lebih lanjut untuk mengetahui seberapa besar pengaruh dari penerapan teknik tersebut terhadap model klasifikasi berita hoaks berbahasa Indonesia dalam meningkatkan kinerja model nantinya.

## 1.2. Identifikasi Masalah

Berdasarkan latar belakang masalah di atas, permasalahan yang dapat diidentifikasi yaitu:

1) Masyarakat memiliki keterbatasan dalam mengidentifikasi berita hoaks secara cepat dan akurat, terutama dalam hal pengetahuan dan rasionalitas
2) Bahasa Indonesia merupakan *low resource language* sehingga sulit mendapatkan kumpulan data besar berbahasa Indonesia
3) Dibutuhkan upaya yang besar untuk membangun kumpulan data besar yang dapat mendukung penelitian
4) Model CNN dan LSTM membutuhkan kumpulan data yang cukup besar untuk dapat menghasilkan kinerja yang baik
5) Melatih model dari awal cukup memakan banyak waktu serta membutuhkan spesifikasi komputer yang tinggi pula

## 1.3. Pembatasan Masalah

Berdasarkan latar belakang dan identifikasi masalah yang telah dijabarkan, batasan masalah yang dapat diidentifikasi yaitu:

1) Model dalam penelitian yang dilakukan akan menggunakan algoritma CNN dan LSTM

2) *Pretrained language model* yang digunakan adalah IndoBERT yang dibuat oleh tim Fajri Koto
3) Data yang digunakan dalam penelitian adalah berita dalam bentuk teks berbahasa Indonesia

### 1.4. Perumusan Masalah

Berdasarkan latar belakang, identifikasi, serta pembatasan masalah di atas maka rumusan masalah dalam penelitian ini adalah "Bagaimana perbandingan hasil kinerja model klasifikasi teks berita hoaks berbahasa Indonesia menggunakan algoritma CNN dan LSTM dengan *transfer learning*?"

### 1.5. Tujuan Penelitian

Berdasarkan latar belakang, identifikasi, batasan, dan perumusan masalah di atas maka tujuan dari penelitian ini adalah:

1) Mengetahui cara mengimplementasikan teknik *deep learning* pada model klasifikasi teks berita hoaks menggunakan algoritma CNN dan LSTM
2) Membandingkan hasil kinerja model klasifikasi berita hoaks berbahasa Indonesia dengan menggunakan algoritma CNN dan LSTM
3) Mengetahui cara mengimplementasikan *pretrained language model* IndoBERT pada model klasifikasi teks berita hoaks berbahasa Indonesia
4) Menerapkan teknik *transfer learning* melalui *pretrained language model* IndoBERT untuk meningkatkan kinerja model
5) Membandingkan hasil kinerja model klasifikasi berita hoaks berbahasa Indonesia dengan dan tanpa *transfer learning*
6) Mengetahui algoritma terbaik berdasarkan tingkat akurasi yang dihasilkan dalam menyelesaikan permasalahan klasifikasi teks berita hoaks berbahasa Indonesia

### 1.6. Manfaat Penelitian

Manfaat yang dihasilkan dari penelitian ini antara lain:

1) Dapat mengetahui model yang sesuai untuk tugas NLP klasifikasi teks dalam kasus klasifikasi teks berita hoaks berbahasa Indonesia
2) Dapat mengetahui pengaruh teknik *transfer learning* dalam meningkatkan kinerja model klasifikasi teks berita hoaks berbahasa Indonesia

3) Dapat dijadikan sebagai dasar acuan dalam penelitian lebih lanjut terhadap pendeteksian teks berita hoaks berbahasa Indonesia menggunakan teknik *deep learning*