BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan zaman modern semakin mempermudah penyebaran informasi melalui internet. Salah satu informasi yang sering tersebar melalui internet dalam bentuk teks adalah berita. Berita merupakan laporan terkait peristiwa-peristiwa yang aktual dan faktual. Namun, tidak semua berita yang sampai pada pembaca bersifat faktual, berita yang tidak faktual adalah berita palsu. Berita palsu disebarkan dari situs web berita maupun pengguna media sosial. Pengguna media sosial yang mengkonsumsi berita tersebut cenderung gagal dalam membedakan berita palsu dengan berita asli [1].

Beberapa penelitian telah mengangkat masalah deteksi berita palsu sebagai topik utama. Salah satunya yaitu penelitian yang telah dilakukan P. Verma dkk. [2] menggunakan metode tradisional yaitu *Term Frequency - Inverse Document Frequency* (TF-IDF) dan *Count Vectorizer* (CV) dibantu dengan *linguistic feature sets* (LFS) untuk *feature extraction*. Lalu *feature* yang sudah dilakukan ekstraksi selanjutnya diklasifikasi menggunakan beberapa algoritma diantaranya yaitu KNN, SVM, Naive Bayes, Decision Tree, Bagging, dan AdaBoost. Dataset yang digunakan merupakan gabungan dari dataset berita di Kaggle, BuzzFeed, McIntire, dan Reuters. Penggabungan dataset ini bermanfaat untuk mengurangi bias dan limitasi individu masing-masing sumber dataset.

Model yang digunakan pada penelitian [2] memiliki hasil yang maksimal dengan akurasi 96.73%. Akurasi terbaik ini dicapai menggunakan algoritma TF-IDF dan CV dengan LFS yang digabungkan secara voting dengan algoritma SVM. Namun, penggunaan metode tradisional seperti TF-IDF dan CV tidak dapat menangkap makna dan konteks dari sebuah kata yang terdapat pada sebuah korpus. Selain itu, metode tersebut tidak dapat melakukan vektorisasi terhadap *out-of-vocabulary terms*.

Selanjutnya, penelitian yang dilakukan oleh Yang S. dkk. [3] menggunakan Bayesian Network untuk *feature extraction* dan pendekatan *collapsed Gibbs sampling* untuk klasifikasi. Bayesian Network digunakan sebagai pengganti *word embedding* untuk *feature extraction* yang bertujuan untuk menangkap hubungan probalitas antar fakta dan opini pada data berita. Lalu klasifikasi dengan pendekatan

collapsed Gibbs sampling berfungsi untuk mengestimasi keaslian berita dan kredibilitas penulis dari data yang bersifat probabilitik. Dataset yang digunakan adalah data *tweet* pengguna Twitter yang pada umumnya bias dan kurang informasi kontekstual.

Model yang diusulkan pada penelitian [3] sangat cocok untuk teks berita yang berasal dari media sosial karena menggunakan Bayesian Network ditahap *feature extraction*. Metode *feature extraction* yang digunakan tidak menangkap makna dari teks. Dimana metode ini menangkap hubungan probabilitas antara opini dan fakta dari teks. *Feature extraction* yang diusulkan tidak dapat digunakan pada teks yang tidak mengandung opini dan fakta secara bersamaan seperti teks berita media pers.

Terakhir, penelitian yang dilakukan oleh Z. A. Khan dan Rekha V. [4] menggabungkan TF-IDF dan Word2Vec untuk *feature extraction* dan diklasifikasikan menggunakan penggabungan *Random Forest* dan *Logistic Regression* dengan *hard voting*. Penggabungkan TF-IDF dan Word2Vec dilakukan dengan cara mengalikan hasil vektorisasi akar kata yang sama. Hal ini bertujuan agar model mampu menangkap pola kemunculan kata pada korpus dan menafsirkan makna dan konteks dari kata tersebut sehingga mampu meningkatkan performa algoritma klasifikasi. Fitur yang sudah diekstraksi akan diklasifikasikan dengan *Random Forest* dan *Logistic Regression* lalu akan voting secara mayoritas (*hard voting*). Model ini mampu mengklasifikasikan berita asli atau palsu dengan skor akurasi 94.24%.

Penggabungan hasil vektorisasi kata dan *word embedding* dengan perkalian (*weighted multiplication*) berpotensi mengurangi kemampuan deteksi pola kata dari TF-IDF dan penafsiran makna kata dari Word2Vec. Metode terbaik untuk menangkap kedua kekuatan dari masing-masing metode vektorisasi kata adalah *weighted average*. Metode penggabungan vektorisasi dengan cara lain kurang efisien, karena frekuensi kata dapat memengaruhi makna pada sebuah model *word embedding* [5]. Selain itu, Menggabungkan beberapa algoritma untuk klasifikasi dengan metode *hard voting* dapat menyebabkan bias saat pengambilan keputusan. *Hard voting* juga kurang efisien digunakan untuk menggabungkan algoritma dalam jumlah genap yang menggabungkan 2 algoritma klasifikasi [4].

Oleh karena itu penelitian yang akan dilakukan memanfaatkan *pretrained word embedding* yaitu *Global Vectors* (GloVe) sebagai *feature extraction*. GloVe adalah model *unsupervised learning* yang sudah dilatih dengan data dari situs-situs web. GloVe mampu menafsirkan makna suatu kata berdasarkan probabilitas

kemunculannya pada suatu korpus secara global. GloVe dapat menafsirkan informasi kontekstual secara lokal berdasarkan pola pada *co-occurrence matrix*. Dalam menangani kata-kata *out of vocabulary* (OOV), GloVe menggunakan *default vectors*[6].

Selanjutnya algoritma klasifikasi berita palsu dan berita asli akan menggunakan *support vector machine* (SVM). SVM dipilih sebagai algoritma klasifikasi karena pada penelitian terkait, nilai akurasi SVM relatif lebih tinggi dibandingkan dengan AdaBoost, KNN, Naive Bayes, *Decision Tree*, dan *Bagging* [2]. Selain itu SVM merupakan model yang fleksibel untuk memproses data dengan kompleksitas yang tinggi seperti data teks. Hasil dari *feature extraction* memiliki dimensi yang tinggi, namun hal ini dapat diatasi oleh SVM dengan menyesuaikan *kernel* yang dipakai. SVM mampu memaksimalkan margin antar kelas dan memiliki ketahanan terhadap data *noise* sehingga tidak menimbulkan *overfitting* [7].

Dari hal-hal yang sudah dipertimbangkan, maka diputuskan untuk memanfaatkan algoritma GloVe untuk *feature extraction* dan algoritma SVM untuk klasifikasi dalam mendeteksi berita palsu. Nilai akurasi, presisi, *recall*, dan F1-score akan digunakan sebagai acuan evaluasi model. Hasil evaluasi akan digunakan untuk menentukan keberhasilan model dalam mendeteksi berita palsu.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah, dapat dirumuskan dua rumusan masalah yaitu;

1. Bagaimana mengimplementasikan *pretrained word embedding* GloVe dan *Support Vector Machine* (SVM) dalam mendeteksi berita palsu?
2. Berapa nilai akurasi, presisi, *recall*, dan F1-score dari hasil klasifikasi model GloVe dan *Support Vector Machine* (SVM) dalam mendeteksi berita palsu?

1.3 Batasan Permasalahan

Dalam melakukan penelitian telah ditentukan beberapa batasan masalah, antara lain:

1. Dataset bersumber dari IEEE Transactions on Computational Social Systems [2] yang merupakan gabungan dari empat dataset penyiaran berita yaitu

Kaggle, BuzzFeed, McIntire, dan Reuters.

2. Data yang digunakan sebagai input berbentuk teks berita dalam bahasa Inggris.
3. Topik berita pada dataset adalah politik Amerika Serikat.

1.4 Tujuan Penelitian

Dari rumusan masalah yang telah dirumuskan didapatkan tujuan penelitian untuk menjawab permasalahan tersebut sebagai berikut:

1. Mengimplementasi *pretrained word embedding* GloVe dan *Support Vector Machine* (SVM) dalam mendeteksi berita palsu.
2. Mengukur nilai akurasi, presisi, *recall*, dan F1-score dari hasil klasifikasi model GloVe dan *Support Vector Machine* (SVM) dalam mendeteksi berita palsu.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat bagi berbagai pihak, berikut manfaat yang dapat diambil yaitu:

1. Memberikan informasi kepada pembaca dalam hal implementasi *machine learning* untuk mendeteksi berita palsu.
2. Mengetahui performa model dalam mendeteksi berita palsu dalam bentuk nilai akurasi, presisi, *recall*, dan F1-score untuk acuan penelitian kedepannya.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN

  Bab 1 merupakan pendahuluan dari penelitian yang telah dilakukan. Bagian ini berisikan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

- Bab 2 LANDASAN TEORI
  Bab 2 membahas landasan teori yang digunakan pada penelitian ini. Teori-teori yang dicantumkan meliputi Berita Palsu, Word Embedding Algorithm, dan Support Vector Machine.

- Bab 3 METODOLOGI PENELITIAN
  Bab 3 berisikan metodologi penelitian yang digunakan dari awal hingga akhir penelitian dan juga perancangan sistem disertai dengan *flowchart*.

- Bab 4 HASIL DAN DISKUSI
  Bab 4 merupakan inti bahasan dan diskusi hasil penelitian yang telah dilakukan.

- Bab 5 KESIMPULAN DAN SARAN
  Bab 5 adalah simpulan dan saran dari penulis atas penelitian yang telah dilakukan.