1. Pendahuluan

Berita atau informasi hoax menurut survey yang didapat dapat didefinisikan dengan berita bohong yang disengaja, berita yang menghasut, atau berita yang tidak akurat. Motif dari berita atau informasi hoax tersebut salah satunya bisa dapat menjatuhkan politik dan bisa mendapatkan dampak yang burut seperti hilangnya reputasi [1][2].

Penelitian ini dilakukan untuk mengatasi permasalahan penyebaran hoax dengan cara mendeteksi berita hoax menggunakan teks mining karena secara otomatis mendeteksi jika terdapat berita hoax. Pengembangan untuk permasalahan mendeteksi berita hoax ini sudah dilakukan beberapa peneliti [3]. Hoax Web Detection menggunakan Support Vector Machine (SVM) berbasis linear kernel yang ditulis oleh Abdillah [4] yang menghasilkan akurasi sebesar 85%.

Klasifikasi yg dilakukan oleh Inggrid [5] masih terdapat kekurangan yaitu hanya bisa melakukan klasifikasi pada 1 bahasa dikarenakan data yang dipakai adalah berita Bahasa Indonesia dan karakteristik berita Bahasa Indonesia berbeda dengan bahasa lainnya. Penelitian lainnya yang dilakukan oleh Agung [6] juga melakukan klasifikasi berita hoax pada 1 bahasa yaitu Bahasa Indonesia. Dengan tren globalisasi terdapat banyak dokumen yang ditulis dengan bahasa yang berbeda-beda dan dengan internasionalisasi komunikasi informasi semakin banyak lembaga bisnis yang melakukan kegiatan internasional. Contohnya website *e-commerce* membutuhkan klasifikasi dan merekomendasikan barang pada bahasa yang berbeda-beda, sehingga klasifikasi yang dapat secara otomatis mengklasifikasikan dokumen dengan bahasa yang berbeda sangat penting [7][8].

Salah satu cara untuk mengatasi kekurangan yang dilakukan oleh [4] adalah dengan metode translasi dataset yang digunakan oleh [9] yang meneliti klasifikasi cross language text classification dengan Bahasa Inggris dan Bahasa Czech yaitu dengan menterjemahkan semua fitur dengan Bahasa Czech menjadi Bahasa Inggris lalu diklasifikasikan. Penelitian ini dilakukan oleh Olsson [9] dikarenakan tidak adanya data train pada datasetnya dan kemungkinan untuk memanfaatkan distribusi topik yang berbeda dalam bahasa yang berbeda untuk meningkatkan klasifikasi keseluruhan untuk pengambilan informasi.

Selanjutnya Agung [6] berkontribusi untuk mendeteksi berita hoax menggunakan pendekatan algoritma klasifikasi Support Vector Machine (SVM) dan Stochastic Gradient Descent (SGD) menggunakan modified huber yang menghasilkan akurasi tertinggi 86%. Sedangkan Santoso [10] mengusulkan sebuah sistem yang menggunakan sosial media Application Program Interface (API) seperti facebook yang digunakan untuk mencari berita hoax pada portal berita lainnya atau sosial media untuk mengecek keaslian dari kiriman berita seseorang.

Berdasarkan penelitian Agung [6] dan Santoso [10] yang telah dilakukan, diketahui bahwa klasifikasi berita hoax terdapat kekurangan yang berkaitan dengan banyaknya pattern fitur atau besarnya dimensi data. Sehingga Agung [6] menyarankan dibutuhkan pemilihan features yang memiliki pengaruh besar dan merupakan sebuah ciri-ciri dari suatu class. Beberapa teknik yang dapat digunakan untuk melakukan pemilihan feature yaitu Gini Index (GI), Information Gain (IG), Mutual Information (MI), Chi-Square (X^2 Statistik) [11].

Data berdimensi besar pada penelitian ini menunjuk pada banyaknya fitur yang digunakan. Pengurangan dimensi merupakan transformasi data dari berdimensi besar menjadi data berdimensi kecil yang mempertahankan beberapa data yang berguna dari data aslinya. Pengurangan dimensi pada biasanya dilakukan jika data berdimensi besar seperti teks diklasifikasi [12]. Beberapa cara dapat dilakukan untuk mengurangi dimensi data salah satunya adalah dengan menggunakan feature selection [12].

Berdasarkan eksperimen yang dilakukan oleh Changqiu [13] untuk mempelajari feature selection kategorisasi teks pada finance menggunakan algoritma klasifikasi NBC, SVM, dan VSM. Feature Selection X^2 Statistik merupakan yang paling efektif karena dapat menghapus banyak fitur sampai dengan 90% tanpa mengurangi akurasi. Sehingga Feature Selection yang akan digunakan pada penelitian ini adalah Mutual Information dan Chi Square.

Kontribusi dari penelitian ini adalah untuk mendeteksi berita hoax “multilingual” yaitu berita dalam bahasa Indonesia, Inggris, dan Spanyol dengan menggunakan metode Machine Translation yang sudah pernah dilakukan oleh [9] dan menggunakan algoritma Support Vector Machine. Penelitian ini juga memanfaatkan Feature Selection yang dapat melakukan klasifikasi berita hoax dengan pattern atau data berdimensi besar untuk melihat jika menghasilkan akurasi yang lebih baik.