1. Pendahuluan

Teknologi telah banyak membawa perubahan, salah satunya teknologi informasi yang menjadi bagian tidak terpisahkan dari kehidupan manusia, seolah melekat seperti pakaian. Dengan adanya teknologi, kehidupan manusia sangat terbantu dalam berbagai hal seperti berkomunikasi jarak jauh, mencari informasi yang tidak bisa dijangkau hingga terciptanya teknologi baru. Teknologi informasi merupakan sebuah bantuan yang diciptakan untuk membantu, mengubah, menyimpan, berkomunikasi dan menyebarkan informasi secara luas ke khalayak umum. Sedangkan teknologi komunikasi merupakan sebuah alat bantu untuk memproses dan mengirim data dari satu perangkat ke perangkat lain [1]. Selain menggunakan koran dan televisi, kini penyebaran informasi bisa melalui situs web (*website*) yang bisa diakses pada media desktop dan handphone kapan pun dan dimana pun.

Berita merupakan sebuah informasi atau kabar yang berhubungan dengan fakta dan sedang terjadi untuk kemudian disampaikan kepada masyarakat [2]. Berita yang termuat dalam situs web (*website*) dinamakan artikel. Banyak media seperti kompas, kumparan, detik.com, BBC Word beralih tempat dari koran ke situs web agar masyarakat bisa mengakses dan mengetahui informasi terbaru. Ada banyak jenis kategori pada artikel yang dimuat, meliputi berita kesehatan, politik, olahraga, hiburan, hingga teknologi terbaru.

Dalam beberapa tahun terakhir, banyaknya artikel yang diunggah menyebabkan masalah pada pengkategorian topik berita, sehingga diperlukannya pengklasifikasian topik berita agar tidak menyebabkan *overload* informasi untuk kemudian pembaca mudah mengakses berita [3]. Klasifikasi dilakukan manual dengan data yang besar membutuhkan waktu lama serta memiliki risiko *human error*, pengguna perlu memikirkan suatu artikel untuk dimasukkan pada kategori tertentu [3]. Maka dari itu perlu dilakukan pengklasifikasian otomatis untuk mengorganisasikan informasi berupa artikel tadi dan pencari informasi bisa paham berdasarkan kategorinya, serta bisa mempermudah pengolahan dan penggunaannya sesuai kebutuhan yang diinginkan.

Klasifikasi merupakan sebuah proses membangun suatu model yang mengkategorikan suatu objek sesuai dengan kelompoknya [4]. Tujuan klasifikasi yaitu mengelompokkan data yang mana data tersebut memiliki target ataupun kelas yang telah ditentukan. Penelitian yang melakukan klasifikasi menggunakan berita online dilakukan oleh siti [1] dengan data sebanyak 500 artikel berita. Penelitiannya membandingkan metode *Support Vector Machine* (SVM) dan *K-Nearest Neigbour* (KNN), perolehan akurasi SVM kernel *Polynomial* mendapat 93.2% dan akurasi KNN mendapat 60%. Dari hasil akurasi tersebut, SVM memiliki akurasi yang tinggi dalam klasifikasi.

Dalam proses klasifikasi mengenal istilah *fitur extraction*, *fitur extraction* merupakan proses dasar dalam kategorisasi yang penting untuk diketahui. Adapun pada penelitian [1] memiliki kekurangan yang tidak melibatkan proses *fitur extraction* tersebut. Fitur penting diketahui karena menjadi proses dasar yang mencerminkan informasi menganai konten dan konteknya [5]. Dalam *machine learning* suatu fitur akan direpresentasikan kedalam bentuk vektor, nilai vektor tersebut diperoleh dari pembobotan *term* (kata) [6]. Penggunaan *fitur extraction* yang umum digunakan yaitu Unigram, Bigram dan Trigram.

Term Frequency - Inverse Document Frequency (TF-IDF) bertujuan untuk memberikan hubungan bobot *term* (kata) berdasarkan frequensi dokumen [7]. *Term Frequency* menunjukkan banyakanya kata pada dokumen dalam satu kalimat dan *Invers Document Frequency* menjelaskan proses untuk menghitung penyebaran *term* dalam dokumen tersebut sehingga menghasilkan data *vector* berupa matriks. Penggunaan TF-IDF dapat disesuaikan dengan fitur N-gram pada bentuk kata yang dikombinasikan dengan metode *machine learning*. Selain pembobotan TF.IDF pada [1] penelitian lain telah melakukan pembobotan berbeda, penelitian [8] melakukan perbandingan pembobotan TF.ABS dan TF.CHI menggunakan metode *Support Vector Machine* (SVM). Dari hasil penelitiannya diperoleh TF.ABS dan TF.CHI memiliki hasil yang sama-sama baik dengan perolehan akurasi yang sama yaitu 95.87%. Penelitian lainnya [9] membandingkan TF.ABS, TF.CHI2, TF.RF, dan TF.IDF menggunakan sebanyak 360 data dengan tujuan penelitian yaitu untuk mengetahui akurasi optimal dikombinasikan dengan *Decision Tree*. Dari hasil penelitian tersebut akurasi tertinggi diperoleh menggunakan pembobotan TF.ABS sebesar 82.22% dibandingkan TF.CHI2 mendapat akurasi sebesar 80.83%, TF.RF 65.56%, dan TF.IDF 50.56%. Metode TF.ABS melakukan kinerja pembobotan dengan melihat kemunculan kata dan kemungkinan kata yang tidak muncul dalam dokumen [8]**.** Adapun pada penelitian ini pembobotan TF-ABS dipilih sebagai pembanding karena pada penelitian [8][9] untuk meningkatkan pembobotan mendapat kesimpulan bahwa TF-ABS lebih baik daripada metode pembobotan lainnya.

Support Vector Machine banyak digunakan untuk mengklasfikasikan berita dalam klasifikasi *text*, pada penelitian [10] mambandingkan tiga metode berbeda yaitu SVM, *Neural Network* (NN) dan *Naive Bayes* (NB). Klasifikasi mengenai berita Nepali (Nepal) ini menggunakan sebanyak 4.964 data. Metode SVM mendapat klasifikasi yang unggul, ini ditunjukkan dalam perolehan akurasi yaitu SVM Kernel RBF mendapat akurasi sebesar74.65%, SVM Linier mendapat 74,62%, NN mendapat 72.99% dan NB hanya mendapat 68.31%. Penelitian lain [11] melakukan klasifikasi *automatic* multilabel untuk artikel berbahasa Indonesia, penelitiannya berfokus mengkomparasikan beberapa seleksi fitur, pembobotan fitur, pendekatan multilabel dan metode klasifikasi, dengan tujuan klasifikasi untuk mengurangi dimensi fitur. Dari hasil percobaan 10 *cross validation* didapat SVM

memperoleh akurasi 85.13%. Berdasarkan penelitian [10] [11] metode SVM mendapat klasifikasi terbaik karena dapat diimplementasikan pada data berdimensi tinggi seperti teks dan volume yang besar kemudian bisa mengurangi dimensi fitur.

Oleh karena itu, penelitian ini bertujuan membangun model klasifikasi untuk menentukan kategori artikel berita online berdasarkan metode pembobotan *Term Frequency - Inverse Document Frequency* (TF.IDF) dan *Term Frequency Absolute* (TF.ABS) dengan algoritma *Support Vector Machine* (SVM). Pendekatan fitur yang digunakan adalah Unigram. Adapun evaluasi dari penelitian ini menggunakan *confusion matrix*. Batasan pada penelitian ini adalah menggunakan tiga buah kernel SVM yakni kernel *linear, polynomial* dna RBF. Jumlah data sebanyak 2225 artikel berita dengan kategori 5 topik yaitu *business, entertainment, politics, sport* dan *tech.*

Bagian selanjutnya membahas studi kasus yang mendukung litelatur penelitian ini. Kemudian bagian sistem yang dibangun serta gambaran penelitian yang disertai penjelasan dari setiap prosesnya. Setelah dilakukan implementasi hasil akan dievaluasi pada bagian evaluasi disertai dengan penjelasan. Bagian terakhir yaitu kesimpulan untuk mendapatkan hasil akhir dari penelitian ini.