

## BAB II

### LANDASAN TEORI

#### 2.1 Cryptocurrency

Kata “*cryptocurrency*” berasal dari gabungan 2 kata yaitu “*cryptography*” yang berarti kode rahasia dan “*currency*” yang berarti mata uang [13]. *Cryptocurrency* adalah sebuah sistem mata uang yang dibuat dengan menggunakan teknologi *blockchain* [2]. Dengan menggunakan teknologi *blockchain* maka setiap data yang ada akan saling terhubung satu sama lain dalam suatu lingkungan pengguna sistem *cryptocurrency* tersebut [2].

Dengan menggunakan *blockchain* maka semua pengguna sistem akan saling terhubung tanpa melalui pihak ketiga, hal ini membuat sistem transaksi menjadi lebih transparan [2]. *Cryptocurrency* tidak memiliki bentuk fisik melainkan hanya sebuah *block data* yang diikat oleh *hash* sebagai validasinya [13]. *Hash* merupakan suatu rangkaian angka dan huruf yang berfungsi untuk memverifikasi validitas informasi [2].

*Cryptocurrency* yang menggunakan *blockchain* sebagai penghubung antar server merupakan teknologi terdesentralisasi yang dapat digunakan secara *peer-to-peer* [14]. Setiap transaksi yang terjadi dalam *cryptocurrency* tercatat dalam sebuah buku besar (*ledger*) yang tidak dapat diubah tanpa persetujuan dari mayoritas pengguna server dalam jaringan. Jaringan tersebut disebut *blockchain*. *Blockchain* sendiri terdiri dari *blockchain public* dan *blockchain private* [15].

#### 2.2 Analisis Sentimen

Analisis sentimen adalah sebuah cara untuk mengetahui pendapat atau kecenderungan seseorang terhadap suatu hal. Tujuan dari analisis sentimen adalah mengetahui sikap atau pendapat dari seseorang terhadap suatu topik atau target. Analisis sentimen dapat digunakan untuk menentukan nilai kesukaan terhadap suatu barang. Nilai tersebut dapat

berupa positif maupun negatif dan hal tersebut dapat dijadikan parameter dalam pengambilan keputusan [16].

Analisis sentimen dapat dilakukan dengan tiga (3) jenis teknik pendekatan yaitu:

a. *Machine learning*

Pendekatan *machine learning* didasarkan pada pembangun klasifikasi dari contoh hasil data yang telah dilabel [17]. Pendekatan *machine learning* memiliki dua (2) jenis teknik yaitu *supervised learning* dan *unsupervised learning*. Pada *supervised learning* pembelajaran didasarkan pada *automatic text classification* dan model klasifikasi dibangun untuk memprediksi kelas berdasarkan kategori yang telah ditentukan [17]. Sedangkan pada *unsupervised learning* pembelajaran tidak bergantung pada domain dan topik data latih, pendekatan ini melakukan pendekatan dengan mengatasi kesulitan mengumpulkan dan pembuatan *data training* yang telah di label [17].

b. *Lexicon-based*

Pada pendekatan *lexicon-based*, analisis sentimen dilakukan dengan menggunakan kata dan frasa opini tanpa mengetahui kata apa yang mengandung opini disusun dan dikumpulkan [17]. Kata – kata dalam teks dievaluasi berdasarkan opini *lexicon* untuk memutuskan orientasi dan sentimen dari teks tersebut [17]. Pembuatan opini *lexicon* dapat dilakukan dengan tiga (3) pendekatan yaitu *manual approach*, *dictionary-based approach*, dan *corpus-based approach* [17]. Pada *manual approach*, kata – kata opini dikumpulkan secara manual berdasarkan pengetahuan dan pengertian masing – masing domain. Pendekatan ini menggunakan waktu yang cukup banyak dan untuk mengatasi hal ini biasanya pendekatan ini dikombinasikan dengan pendekatan *automated* agar memperbaiki kesalahan yang dibuat oleh pendekatan *manual* [17].

Pada *dictionary-based approach*, kata – kata opini didapatkan dari sumber *lexicographical* seperti kamus *online*. Pada pendekatan ini digunakan sinonim, antonim, dan hierarki dalam opini *lexicon* untuk menentukan sentimen kata. Pendekatan ini memiliki keterbatasan dalam sentimen dengan konteks yang spesifik karena tidak ada pengetahuan khusus tentang suatu konteks. Kamus yang sering digunakan adalah *WordNet*, *SentiWordNet*, *secticNet*, *sentifull*, dll [17]. Sedangkan pada *corpus-based approach*, pendekatan ini memanfaatkan pola sintaksis kata – kata yang muncul bersamaan dengan kata – kata opini untuk mengidentifikasi dan menyusun kata – kata opini dalam *corpus* besar [17]. Pendekatan ini menghilangkan batasan konteks yang spesifik dalam klasifikasi kata – kata opini yang dimiliki oleh *dictionary-based approach*.

c. *Hybrid approach*

*Hybrid approach* merupakan kombinasi antara pendekatan *machine learning* dan *lexicon-based* [17]. Pendekatan ini memiliki keuntungan yaitu dapat membuat deteksi dan pengukuran sentimen pada tingkat konsep dan akurasi yang tinggi dari sebuah algoritma *supervised learning*. Terdapat banyak penelitian yang dapat membuktikan bahwa dengan mengombinasikan kedua pendekatan ini akan memberikan peningkatan kinerja pada klasifikasi [17].

## 2.3 Data Preprocessing

merupakan langkah yang dilakukan untuk mengubah data mentah menjadi sebuah data yang memiliki format terstruktur dan dapat dimengerti [5]. *Data preprocessing* dilakukan agar dapat mengurangi dimensi data tanpa mempengaruhi klasifikasi [18]. Dalam *data preprocessing* terdapat beberapa langkah yang dilakukan untuk menghilangkan dan mengatasi *noisy data*, berikut adalah tahapannya:

a. *Case Folding*

*Case folding* merupakan proses penyamaan seluruh teks (*case*) pada suatu kalimat menjadi huruf kecil. *Case folding* dilakukan karena tidak semua teks dalam suatu kalimat memiliki konsistensi dalam penggunaan huruf kapital [19]. Contoh dari *case folding* adalah “hArI iNi” menjadi “hari ini”.

b. *Data Cleaning*

*Data cleaning* merupakan proses penghapusan karakter – karakter lain selain alfabet a – z seperti tanda baca, URL atau *link*, *hashtag*, dan *username* [20]. Contoh dari *data cleaning* adalah “halo @xdyt21! Kunjungi website kami di [www.bSddsf.com](http://www.bSddsf.com) untuk mendapatkan #diskon ya” menjadi “halo xdyts kunjungi website kami di untuk mendapatkan diskon ya”.

c. *Stemming*

*Stemming* merupakan proses untuk mengubah kata – kata dalam suatu kalimat menjadi kata dasar atau kata akar (*root word*) [20]. Pada proses ini terdapat dua pendekatan yaitu dengan pendekatan kamus dan pendekatan aturan [21]. Contoh dari *stemming* adalah “menyapu lantai” menjadi “sapu lantai”.

d. *Stopword Removal*

*Stopword removal* merupakan proses mengurangi jumlah kata yang tidak memiliki arti penting dan tidak digunakan [21]. Tujuan dari *stopword removal* adalah mengoptimalkan kinerja klasifikasi sekaligus mengurangi data *sparsial* serta menyusutkan ruang fitur secara *substantial* [20].

e. *Tokenisasi*

*Tokenisasi* merupakan proses pemisahan setiap kata dalam suatu kalimat dalam suatu dokumen [20]. Pada proses ini dilakukan pemotongan dokumen menjadi pecahan kecil yang dapat berupa bab, sub-bab, paragraf, kalimat, dan kata (*token*). Tujuan dari proses ini adalah menghilangkan *whitespace* [21].

f. *Lemmatization*

*Lemmatization* merupakan proses mentransformasi kata – kata yang muncul dalam teks menjadi bentuk dasar (*lemma*). Pada proses ini dilakukan normalisasi di mana berbagai varian kata yang berbeda dari sebuah kata dipetakan ke *lemma* yang sama sehingga bisa dianalisis sebagai satu item (istilah atau konsep) [22].

## 2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

*TF-IDF* adalah sebuah metode yang dapat menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval* [19]. *TF-IDF* bekerja dengan membobotkan setiap kata pada sebuah kalimat berdasarkan *TF* dan *IDF* [18]. *TF* (*Term Frequency*) merupakan pembobotan yang dilakukan berdasarkan jumlah kemunculan sebuah kata pada kalimat [18]. Sedangkan *IDF* (*Inverse Document Frequency*) merupakan sebuah cara untuk membobotkan kepentingan sebuah kata [18]. Adapun rumus yang digunakan dalam perhitungan *TF-IDF* adalah sebagai berikut [23]:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \quad (2.1)$$

$$IDF_t = \log \frac{N}{DF_t} \quad (2.2)$$

Dimana:

$t$  = kata – kata yang dihitung

$d$  = bobot kalimat ( $d$ )

$TF-IDF_{t,d}$  = kalimat bobot ( $d$ ) terhadap kata ( $t$ )

$TF_{t,d}$  = *Term Frequency*

$IDF_t$  = *Inverse Document Frequency*

$N$  = Jumlah kalimat

$DF_t$  = Jumlah kata yang diulang

## 2.5 Chi-square

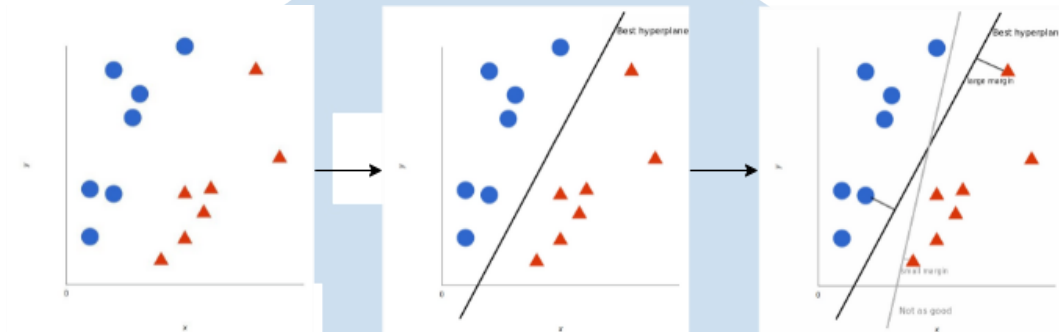
*Chi-square* merupakan sebuah metode yang dapat mengukur hubungan antar dua (2) variabel [24]. Teknik ini mengukur distribusi nilai ketergantungan antara fitur dan kelas [25]. Bila *value* dari sebuah variabel adalah nol (0) berarti variabel tersebut merupakan variabel independen. Dan apabila variabel tersebut memiliki *value* lebih besar daripada  $X^2$ , berarti variabel tersebut memiliki hubungan satu sama lain. Pada analisis sentimen, *Chi-square* digunakan sebagai teknik fitur seleksi [24]. Adapun rumus yang digunakan dalam perhitungan *Chi-square* adalah sebagai berikut [24] :

$$\tilde{X}^2(f) = \sum_{k=1}^c \frac{n_k}{N} \times X^2(f, c_k) \quad (2.3)$$

## 2.6 Support Vector Machine (SVM)

*Support Vector Machine (SVM)* adalah metode yang diperkenalkan oleh Vapnik pada tahun 1992 [26]. *Support Vector machine* merupakan salah satu algoritma yang termasuk dalam *supervised learning* yang dapat digunakan untuk proses klasifikasi dan regresi [20]. *Support Vector Machine* memiliki fungsi untuk memisahkan kumpulan data menjadi dua kelas yang berbeda dengan menggunakan *hyperlanes* sebagai batas keputusan [11]. Dalam mengambil keputusan *Support Vector Machine* menggunakan fungsi kernel yang berfungsi untuk membantu dalam memetakan data [11]. Kernel yang biasanya digunakan adalah *polynomial*,

*sigmoid*, *Radial Basis Function* [11]. Gambaran cara kerja *Support Vector Machine* dijelaskan pada gambar 2.1 berikut:



Gambar 2.1 Cara Kerja *Support Vector Machine*

## 2.7 Evaluation

### 2.7.1 Accuracy

*Accuracy* merupakan perbandingan antara data yang terklarifikasi benar dengan keseluruhan data [27]. Pada umumnya nilai *accuracy* merupakan representasi dari model yang memiliki kinerja yang baik. *Accuracy* dapat dihitung dengan menggunakan rumus sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2.4)$$

### 2.7.2 Recall

*Recall* merupakan nilai yang didapatkan dari kategori positif yang terklasifikasi secara benar oleh model [27]. *Recall* dapat dihitung dengan menggunakan rumus:

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2.5)$$

### 2.7.3 Precision

*Precision* merupakan gambaran jumlah data kategori positif yang diklasifikasi secara benar dengan total data yang diklasifikasi positif [28]. *Precision* dapat diperoleh dengan perhitungan:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (2.6)$$

### 2.7.4 F1-Score

*F1-Score* merupakan rata – rata dari *precision* dan *recall*. Pada *F1-score*, 1 diartikan sebagai nilai maksimum dan 0 adalah nilai minimum [28]. *F1-Score* dapat dihitung dengan menggunakan perhitungan:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.7)$$

### 2.7.5 Confusion Matrix

*Confusion matrix* merupakan metode yang digunakan untuk meninjau hasil klasifikasi yang diperoleh dengan cara menghitung frekuensi kebenaran yaitu dengan menghitung *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN). *TN* merupakan nilai yang didapatkan ketika model dapat mengklasifikasi data negatif menjadi negatif, *FP* merupakan nilai yang didapatkan ketika model mengklasifikasi data negatif menjadi positif, *FN* merupakan nilai yang didapatkan ketika data positif diklasifikasi menjadi negatif, dan *TP* merupakan nilai yang didapatkan ketika data positif diklasifikasi menjadi positif. [28].

## 2.8 Penelitian Terdahulu

Adapun penelitian sebelumnya terkait dengan *cryptocurrency* yang dijadikan acuan pada penelitian ini yaitu Tabel 2.1.

Tabel 2.1 Penelitian Terdahulu

| Nama Jurnal                              | Nama Artikel<br>(Vol, Tahun)                                                                                                    | Nama Penulis                                           | Masalah                                                                                                                                                | Metode                                                                                                                                                                                                              | Hasil                                                                                                                                                                                                                                                                                                     |
|------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>International Journal on ICT</i>      | “Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection”( Vol. 7, 2021)[25] | (N. S. Azzahra, D. T. Murdiansyah, and K. M. Lhaksana) | Media sosial seringkali disalahgunakan sebagai sarana penyebaran hal – hal tidak baik seperti kebencian, komentar rasis, radikalisme, pornografi, dll. | Melakukan klasifikasi komentar negatif dengan menggunakan SVM dengan tiga jenis kernel berbeda dan TF-IDF sebagai feature extraction serta Chi-square sebagai feature selection                                     | Dari semua <i>kernel</i> , dengan menggunakan SVM kernel <i>linear</i> , TF-IDF sebagai <i>feature extraction</i> dan <i>Chi-square</i> sebagai <i>feature selection</i> , percobaan tersebut dapat menghasilkan <i>F1-score</i> sebesar 76,55%                                                           |
| <i>J. King Saud University</i>           | “Feature selection using an improved Chi-square for Arabic text classification”( Vol. 32, 2020)[29]                             | (S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi)   | Sulitnya klasifikasi pada dokumen aksara Arab karena keterbatasan dalam sumber dokumentasi                                                             | Melakukan test perbandingan <i>classifier</i> ( <i>Decision Tree</i> dan SVM) dan <i>feature selection</i> ( <i>Chi-square</i> , <i>Information Gain</i> , <i>Improved Chi-square</i> , <i>Mutual Information</i> ) | Dari semua percobaan antara <i>classifier</i> dan <i>feature selection</i> , <i>classifier SVM</i> dengan <i>feature selection Chi-square</i> memiliki tingkat keberhasilan paling tinggi yakni dengan mencapai nilai <i>precision</i> sebesar 85,29%, <i>recall</i> 85,17%, dan <i>F-measures</i> 84,93% |
| <i>Multimedia Tools and Applications</i> | “Efficient feature selection techniques for sentiment analysis” (Vol. 79, 2019) [24]                                            | (Avinash Madasu, Sivasankar Elango)                    | Cara memilih <i>feature selection</i> yang cocok dengan <i>classifier</i>                                                                              | Melakukan test performa dengan membandingkan tiga jenis <i>dataset</i> , tujuh jenis <i>feature</i>                                                                                                                 | Dari semua <i>dataset</i> , terdapat 1 <i>dataset</i> yang memiliki hasil akurasi terbaik yaitu dengan menggunakan algoritma SVM dan <i>Chi-square</i>                                                                                                                                                    |

|                |                                                                                                                                                          |                                           |                                                                    |                                                                                                                   |                                                                                                                                                                                                                                                                                                                 |
|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------|--------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                |                                                                                                                                                          |                                           |                                                                    | <i>selection</i> , serta enam <i>classifier</i> yang berbeda                                                      | sebagai <i>feature selection</i> sebesar 82%                                                                                                                                                                                                                                                                    |
| ULTIMATICS     | “Implementasi Algoritma <i>Support Vector Machine</i> dan <i>Chi Square</i> untuk Analisis Sentimen <i>User Feedback Aplikasi</i> ” (Vol. 12, 2020) [30] | (L. Luthfiana, J. C. Young, and A. Rusli) | Keterbatasan dalam penggunaan <i>user feedback</i> sebuah aplikasi | Melakukan analisa sentimen rating aplikasi dengan menggunakan <i>Support Vector Machine</i> dan <i>Chi-square</i> | Dengan menggunakan fitur seleksi <i>Chi Square</i> dapat meningkatkan hasil akurasi pada pengklasifikasiannya dengan hasil akurasi 77%, <i>precision</i> 50%, <i>recall</i> 55%, dan <i>F1-Score</i> 73%.                                                                                                       |
| ULTIMA InfoSys | “Apakah <i>Youtuber</i> Indonesia Kena <i>Bully Netizen?</i> ” (Vol. 11, 2020) [31]                                                                      | (J. Siahaan, W. Wella, and R. I. Desanti) | Banyaknya komentar yang ditinggalkan pada akun-akun terkenal       | Menganalisis seberapa sehat atau tidak sehatnya lingkungan internet di Indonesia                                  | Analisa dengan menggunakan model <i>Support Vector Machine (SVM)</i> dapat menghasilkan akurasi sebesar 81.2%, yaitu terdapat 49.524% komentar yang mengandung unsur <i>cyberbullying</i> , yang berarti dapat dikatakan bahwa <i>Youtuber</i> Indonesia tidak selalu dirundung oleh para masyarakat Indonesia. |

Pada Tabel 2.1 terdapat lima penelitian pendahulu yang digunakan sebagai acuan dalam penelitian ini. Jurnal pertama yang berjudul “*Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection*” merupakan jurnal penelitian dimana dilakukan perbandingan algoritma SVM dengan berbagai *kernel* dan *Chi-square* dalam beberapa proporsi untuk mencari hasil terbaik dalam mengklasifikasi komentar negatif. Dari tiga kernel yang dibandingkan yaitu *linear*, *sigmoid*, dan *RBF* serta proporsi *k* pada *Chi-square* yaitu 20%, 40%, 60%, 80%, ditemukan bahwa dengan menggunakan *kernel linear* dan proporsi *k* pada *Chi-square* sebesar 20% didapatkan hasil *F1-Score* terbaik pada angka 76,55% [25]. Pada jurnal kedua yang berjudul “*Feature selection using an improved Chi-square for Arabic text classification*”,

penelitian tersebut meneliti perbandingan antar *classifier Decision Tree* dan *SVM* serta *feature selection Chi-square, Information Gain, Improved Chi-square, Mutual Information* dalam melakukan klasifikasi terhadap dokumen aksara Arab. Dari kedua *classifier* dan empat *feature selection* ditemukan bahwa dengan menggunakan *classifier SVM* dan *feature selection Chi-square* dapat mencapai nilai *precision* sebesar 85,29%, *recall* 85,17%, dan *F-measures* 84,93% [29]. Pada jurnal ketiga yang berjudul “*Efficient feature selection techniques for sentiment analysis*” dilakukan penelitian dengan melakukan *test* performa pada tiga jenis *dataset*, tujuh jenis *feature selection*, serta enam *classifier* yang berbeda dalam pencarian *classifier* yang cocok dengan *feature selection* pada tiga *dataset* yang berbeda. Pada penelitian tersebut ditemukan hasil yaitu dengan membandingkan semua *dataset*, terdapat 1 *dataset* yang memiliki hasil akurasi terbaik yaitu dengan menggunakan algoritma *SVM* dan *Chi-square* sebagai *feature selection* sebesar 82% [24]. Pada jurnal keempat yang berjudul “Implementasi Algoritma *Support Vector Machine* dan *Chi Square* untuk Analisis Sentimen *User Feedback* Aplikasi” dilakukan penelitian dengan menggunakan *user feedback* pada sebuah aplikasi untuk mencari tahu apakah *Chi-square* memiliki pengaruh terhadap performa klasifikasi. Dari kedua skenario yang telah dilakukan pada penelitian tersebut, didapatkan hasil bahwa *Chi-square* dapat mempengaruhi performa klasifikasi. Pada jurnal kelima yang berjudul “Apakah Youtuber Indonesia Kena Bully Netizen?” penelitian tersebut menganalisa komentar di akun sosial media milik Youtuber Indonesia apakah komentar – komentar tersebut merupakan komentar sehat atau tidak dengan menggunakan *Support Vector Machine* untuk melakukan analisis sentimen. Hasil dari analisis tersebut membuahkan hasil bahwa tidak selalu Youtuber Indonesia memiliki komentar tidak sehat. Hal tersebut dibuktikan dengan rata – rata persentase komentar negatif pada sepuluh Youtuber top di Indonesia adalah 49,524%. Berdasarkan jurnal - jurnal tersebut maka dilakukan adopsi untuk penelitian ini yaitu dengan menggunakan algoritma *Support Vector Machine* sebagai

*classifier* serta *Chi-square* sebagai *feature selection* dalam melakukan analisis sentimen. Selain mengadopsi, dilakukan pembaharuan yaitu dengan menggunakan data dengan objek yang berbeda. Data yang digunakan pada penelitian ini merupakan data yang memiliki kata kunci yang berhubungan dengan *cryptocurrency* dan menggunakan dua varian data yaitu *tweet* dan *retweet*.

