# BAB II

# TINJAUAN PUSTAKA

Deteksi plagiarisme intrinsik merupakan sebuah metode untuk mendeteksi plagiarisme dengan cara mengenali gaya tulisan dari penulis. Gaya tulisan ini dapat dihitung berdasarkan jumlah kata yang diulang antar kalimat, tanda baca yang digunakan, kelas kata yang digunakan dan lain-lain. Dengan mengenali gaya penulisan, metode ini dapat mendeteksi adanya plagiarisme dalam penulisan yang bukan berasal dari hasil penjiplakan karya orang lain.

Beberapa penelitian telah dilakukan untuk mengembangkan metode deteksi plagiarisme intrinsik, antara lain penelitian yang dilakukan oleh Stein et al. (2011), menjelaskan bahwa deteksi plagiarisme intrinsik dapat dilakukan dengan menggunakan suatu *threshold* yang digunakan sebagai *outlier* untuk mdenteksi perubahan gaya tulisan dari penulis. Proses yang dilakukan pada algoritme ini meliputi *impurity assesment*, *decomposition strategy*, *style model construction*, *outlier indentification* dan *outlier post-processing*. *Impurity assesment* dilakukan pada dokumen untuk menganalisis tipe dokumen (*paper, dissertation, assingment*), genre (*nover, report, research*), dan juga institusi yang membuatnya. *Decomposition strategy* merupakan sebuah proses dimana dokumen yang telah dianalisis diekstrak setiap fiturnya. Fitur-fitur yang dapat diekstrak dari sebuah dokumen antara lain *lexical features*, *syntatic features* dan *structural features*. Setelah semua fitur diekstrak, fitur-fitur ini dijadikan sebuah vektor pada bagian *style model construction*. Kemudian dicari *outlier* yaitu sebuah bagian dimana kalimat pada sebuah dokumen telah melewati batas atau *treshold* yang ditentukan, yang kemudian outlier ini akan di analisis kembali untuk di-*learning*. Model ini menggunakan 30 *lexical features* untuk digunakan sebagai batas atau *treshold*.

Algoritme untuk menentukan sebuah dokumen memiliki unsur plagiarisme dengan menggunakan *intrinsic plagiarism* memiliki tiga tahap yaitu: menyusun fitur dasar untuk segmen teks, membangun fungsi *author style*, dan *post-processing* dengan *outliers detection*. Untuk merepresentasikan dalam bentuk vektor teks dengan menggunakan *Vektor Space Model* (VSM) dan membuat deskripsi fitur, pemetaan dengan metode n-gram, *average word frequency* dan *vector space model* diperlukan (Kuznetsov et al., 2016). Semakin banyak kata-kata spesifik yang dimiliki sebuah kalimat, semakin menyimpang kalimat tersebut dari gaya penulisnya. *Text parsing*,

dokumen dan setiap kalimat di bagi menjadi karakter-karakter n-grams, dari pada kata. Eksperimen menunjukan bawaha penggunaan 1-grams, 3-grams, dan 4-grams bersamaan memberikan praktek terbaik. Setiap kalimat juga menghitung berapa banyak penggunaan simbol-simbol pada akhir kalimat, dan POS tags menggunakan nltk parser. Untuk memprediksi plagiarisme, algoritma yang digunakan adalah *Gradient Boosting Regression Trees*. Outputnya merupakan *Author Style Function*, sebuah kombinasi fitur model, dan skor sebuah kalimat yang mengindikasi derajat ketidakcocokan dengan gaya penulisan author. Hasil dari penelitian yang dilakukan oleh Kuznetsov et al. (2016) menghasilkan *F1 score* sebesar 32%.

Kesalahan penelitian atau penyalahgunaan teks didefinisikan sebagai salah satu atau kombinasi dari pemalsuan, atau plagiarisme. Istilah-istilah ini dijelaskan sebagai berikut: fabrikasi mengacu pada pembuatan data, pemalsuan mengacu pada distorsi data, dan akhirnya plagiarisme yang umumnya dilakukan oleh akademisi, merupakan sebuah perbuatan menyalin pekerjaan orang lain tanpa rujukan yang tepat. Distorsi data yang berupa pemalsuan data ini mengakibatkan adanya data-data "sampah" yang tidak sesuai dengan keadaan dan membuat penelitian tersebut tidak valid. Untuk mengetahui distorsi data tersebut, dibuatlah sebuah tools yaitu *Text Misuse Detection Tool* (TMDT) yang terdiri dari *Global Similarity Index*, *Moderator*, *Modified Stylometry Method* dan *Latent Semantic Indexing* (Alsallal et al., 2013). *Latent Semantic Analysis* atau LSI adalah pendeteksi kumpulan kata yang tidak tergantung pada bahasa utama yang terlalu banyak terwakili atau mirip manusia (frasa). *Modified Stylometry Method* (MSM) menganalisis gaya penulisan setiap dokumen corpus, membuat *template* untuk frasa mirip manusia yang dikenali. Moderator menerima input dari MSM dan LSI, dan membandingkan "hit" signifikan yang terdeteksi dari yang terakhir terhadap templat untuk *template* yang dikenali dan mirip manusia yang disempurnakan dari yang sebelumnya.

Ketika suatu dokumen dicurigai terdapat plagiarisme di dalamnya dan tidak dapat melakukan perbandingan dengan dokumen lain, perlu dilakukan pengecekan dengan dokumen itu sendiri, dengan menggunakan metode deteksi plagiarisme intrinsik. Metode deteksi plagiarisme dengan n-gram merupakan sebuah metode yang diperkenalkan oleh Bensalem et al. (2014) yang terinspirasi dari Grozea dan Popescu (2010) dalam papernya yang berjudul *plagiarism direction detection*. Dalam penelitian yang dilakukan oleh Grozea dan Popescu (2010) dijelaskan bahwa bagian dari 8-gram lebih banyak terdapat pada dokumen aslinya daripada ada di dokumen yang telah di plagiat. Grozea dan Popescu (2010) menyakini bahwa plagiarisme dalam sua-

tu dokumen bisa di ketahui dari kemunculan jumlah fragmen n-gram dalam dokumen yang mencurigakan. Semakin banyak fragmen n-gram yang jarang ada di dokumen, ada kemungkinan dokumen tersebut merupakan hasil penjiplakan. Dengan menggunakan 6-grams dan 4 fitur pada metode n-gram dibandingkan dengan metode Grozea dan Popescu (2010) menunjukan bahwa ada perbedaan yang signifikan pada metode dengan menggunakan n-gram.

*TextTilling* merupakan sebuah algoritme yang diperkenalkan oleh Hearst (1997), yang berupa sebuah metode untuk mengetahui perubahan sebuah topik berubah. Algoritme ini menjadi salah satu metode untuk menganalisis plagiarisme dalam sebuah dokumen. Dalam algoritme ini, terdapat *tokenization* dan *lexical score determination*. Kedua hal tersebut digunakan untuk mentoken-token dokumen menjadi paragraf dan kalimat, kemudian dihitung *lexical score*-nya untuk mencari perubahan topik di setiap kalimatnya. Hasil peneltian ini, dapat memprediksi perubahan topik sampai dengan 80%.

Penelitian ini akan didasarkan pada penelitian Kuznetsov et al. (2016), dengan penambahan *lexical score*, jumlah kalimat yang digunakan dan jumlah *n-gram* yang digunakan.

**Tabel 2.1: Perbandingan penelitian yang telah ditinjau**

| No | Peneliti | Topik | Metode |
|---|---|---|---|
| 1 | Hearst, 1997 | Melakukan penelitian dengan menggunakan algoritma *Text-Tilling* untuk segmentasi teks | • *Tokenize*<br>• *Lexical Score* |
| 2 | Stein et al., 2011 | Melakukan penelitian dengan menggunakan *Intrinsic Plagiarism* untuk mendeteksi plagiarisme | • *VSM*<br>• *Style Model Contruction*<br>• *Outlier Identification* |
| 3 | Alsallal et al., 2013 | Melakukan penelitian dengan *Latent Semantic Indexing* dan *Stylometry* untuk mendeteksi Plagiarisme | • *Latent Semantic Indexing*<br>• *Stylometric Technique*<br>• *Post Publication Review* |
| 4 | Bensalem et al., 2014 | Melakukan penelitian dengan menggunakan n-gram untuk mendeteksi plagiarsime secara intrinsik | • *n-gram* |
| 5 | Kuznetsov et al., 2016 | Melakukan penelitian dengan *intrinsic plagiarism* dan *author diarization* untuk mendeteksi plagiarisme | • *Stylometry*<br>• *n-gram* |
| 6 | Grozea dan Popescu, 2010 | Melakukan penelitian dengan menghitung jumlah *fragment n-gram* | • *n-gram* |