BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam penelitian ini akan digunakan sepuluh tinjauan studi yang nantinya dapat mendukung penelitian. Berikut ini merupakan tinjauan studi yang berkaitan dengan *Detection Language,* dapat dilihat pada Tabel 2.1 Daftar Literatur.

Tabel 2.1 Daftar Literatur

No. Literatur	Penulis	Informasi Publikasi (Volume, Tahun, ISSN, Penerbit)	Judul
Literatur 01	Tommi Jauhiainen Marco Lui Marcos Zampieri Timothy Baldwin Krister Linden	2018, Journal of Artificial Intelligenece Research 65	Automatic Language Identification in Text: A Survey
Literatur 02	Muhammad Okky Ibrohim	2018, 3rd International Conference on Computer Science and	A Dataset and Preliminatries Study for Abusive

	Indra Budi	Computational Intelligence, Procedia Computer Science 135	Language Detection in Indonesian Social Media
Literatur 03	Marco Lui Jey Han Lau Timothy Baldwin	2014, Computational Linguistics, Q14-1003	Automatic Deteciton and Language Indentification of Multilingual Documents
Literatur 04	Anna Schmidt Michael Wiegand	2017, Association for Computional Linguistics	A Survey on Hate Speech Detection using Natural Language Processing
Literatur 05	Chikashi Nobata Joel Tetreault Achint Thomas Yashar Mehdad Yi Chang	2016, International World Wide Web Conference Comitte(IW3C2)	Abusive Language Detection in Online User Content

Literatur 06	Agus Zainal Arifin Yuita Arum Sari Evy Kamilah Ratnasari Siti Mutrofin	2014, Modern Education and Computer Science (MECS)	Emotion Detection of Tweets in Indonesia Language using Non-Negative Matrix Factorization
Literatur 07	Rasha Hassan Abbas Firas Abdul Elah Abdul Kareem	2019, Journal of Soutwest Jiaotong University, 0258-2724	Text Language Indentification Using Letters (Frequency, Self-Information, and Entropy) Analysis for English, French and German Language
Literatur 08	Edgar Chavesz Moises Garcia Jesus Favela	2015, Research in Computing Science	Fast And Accurate Language Detection in Short Text using Contextual Entropy

Literatur 09	INggrid Yanuar Risca Pratiwi Rosa Andrie Asmara Faisal Rahutomo	2017, International Conference on Information & Communication Technology and System (ICTS)	Study of Hoax News Detection Using Naïve Bayes Classifier in Indonesian Language
Literatur 10	Enrique Flores Alberto Barron- Cedeno Paolo Rosso Lidia Moreno	2011, Dpto. De Sistemas Informaticos y Computacion, Universidad Polinecnice de Valencia	Towards the Detection of Cross- Language Source Code Reuse

2.1.1 Literatur 1

(Jauhiainen, et al., 2018) Penelitian ini berjudul"AUTOMATIC LANGUAGE IDENTIFICATION IN TEXTS A SURVEY"

Hingga saat ini, indetifikasi bahasa menjadi bagian penting dari banyaknya pengolahan teks, sebagai salah satu teknik pengolahan teks secara umum di asumsikan pada bahasa dari teks masukan yang diketahui. Pada penelitian ini menyajikan penelitian dan survei yang luas dari fitur serta metode yang digunakan pada indetifikasi bahasa. Pengenalaan teks pada suatu spesifikasi bahasa muncul secara alami untuk para pembaca dan dapat dikenali. Dan banyak sumber juga menghadirkan pengenalan teks ini pada topik *Natural Language Processing* (NLP) menurut beberapa penulis bahwa pengolahan bahasa alami merupakan bidang ilmu computer, kecerdasan buatan, dan linguistic yang berkaitan dengan interaksi antara komputer dan bahasa alami (Luiet al., 2018). Sedangkan penelitian ini bertujuan untuk dapat mengikuti atau meniru kemampuan seseorang untuk dapat mengenali bahasa tertentu. Kemampuan sistem semacam itu dapat digambarkan sebagai kelebihan *super* seperti pada rata-rata orang yang mungkin dapat mengidentifikasi beberapa bahasa, dan ahli bahasa atau penerjemah yang terlatih. Pada umumnya penelitian *Language Indentification* (LI) berlaku untuk modalitas bahasa apapun, termasuk juga ucapan, dan teks tulisan tangan. Dan semua yang relevan untuk sarana penyimpanan informasi yang melibatkan bahasa, digital, atau lainnya. Namun pada penelitian ini membahas pada survei, yang melibatkan pemberian kode secara digital. Dalam *language identification monolingual*, pengerjaan tersebut dilakukan untuk dapat memberikan label bahasa yang unik pada setiap dokumen.

Namun, untuk dapat mencapai keakuratan tersebut, asumsi yang disederhanakan harus dilakukan layaknya seperti monolingualitas yang disebutkan diatas dari setiap dokumen, dan juga asumsi tentang jenis dan jumlah data serta jumlah bahasa yang digunakan atau dipertimbangkan.

2.1.2 Literatur 2

(Ibrohim & Budi, 2018) Penelitian ini berjudul"A DATASET AND PRELIMINARIES STUDY FOR ABUSIVE LANGUAGE DETECTION IN INDONESIAN SOCIAL MEDIA"

Bahasa atau kata yang kasar adalah suatu ekspresi dapat berupa teks ataupun lisan yang berisi kata-kata atau ungkapan kasar/kotor baik dalam konteks lelucon, kata yang merujuk konteks vulgar untuk seseorang. Dan termasuk juga dimana banyak orang di media sosial membuat suatu tulisan dan postingan dengan bahasa yang kasar, contoh beberapa di antaranya adalah *facebook, line, twitter,* dll. Masalah untuk dapat mendeteksi bahasa-bahasa tersebut pada media sosial dapat dikatakan adalah masalah yang cukup sulit untuk dapat diselesaikan karena masalah tersebut tidak mudah untuk dapat diselesaikan hanya mengandalkan kata yang sama atau cocok. Dalam penelitian ini mendahulukan pembelajaran untuk mendeteksi bahasa yang kasar pada media sosial yang digunakan di Indonesia. Dan juga mencoba untuk menciptakan sebuah media atau system untuk dapat mendeteksi bahasa kasar, khususnya pada media sosial. Serta membangun *machine learning* yang dapat meberikan laporan terkait dengan deteksi tersebut. Metode yang digunakan pada penelitian ini menggunakan *Naïve Bayes*, *Support Vector Machine*, dan *Random Forest*

Decesion Tree classifier untuk menindentifikasi apakah sebuah *tweet*/cuitan pada media sosial tidak termasuk kedalam kata yang kasar.

2.1.3 Literatur 3

(Lui, et al., 2014) Penelitian ini berjudul "AUTOMATIC DETECTION AND LANGUAGE INDENTIFICATION OF MULTILINGUAL DOCUMENTS"

Identifiakasi bahasa dapat dilakukan secara otomatis dengan cara menerapkannya pada sebuah dokumen. Dimana dalam pengerjaannya, sering terjadi kesalahan mendeteksi bahasa yang terkandung dalam banyaknya jenis bahasa yang diinginkan. Oleh karena itu dikenalkanlah sebuah metode yang dapat mendeteksi bahasa secara otomatis dan juga berbagai bahasa untuk suatu dokumen. Dengan mengindentifikasikan bahasa yang ada serta takaran atau jumlah banyaknya bahasa yang ada pada dokumen tersebut dengan sedemikian menjadi rinci. Deteksi bahasa tersebut, dimulai dari teknik NLP (*Natural Language Processing*) awal yang menyajikan *monolingual (*satu bahasa), termasuk juga dalam data bunyi atau pengucapan untuk pengenalan bahasa asing dan juga dapat menurunkan kemampuan dari sistem NLP (Alex, et al., 2007; Cook and Lui, 2012). Deteksi dokumen multibahasa otomatis juga dapat digunakan sebagai sarana tahap penyaringan awal untuk meningkatkan kualitas dari data masukan. Sehingga deteksi tersebut juga penting untuk pencapaian data linguistic pada *web* dan penerapan *mining bilingual text* untuk statistic mesin penerjemah dalam sumber *online*. Dengan pengenalan suatu metode yang dapat mendeteksi dokumen multibahasa dan simulasi berbagai bahasa yang di hadirkan dengan disajikan dalam porsi/persentase pada dokumen tertulis

tersebut. Pencapaian dalam probabilitas yang akan digabungkan, menggunakan representasi pengembangan dokumen untuk dapat di indetifikasi.

2.1.4 Literatur 4

(Schmidt & Wiegand, 2017) Penelitian ini berjudul "A SURVEY ON HATE SPEECH DETECTION USING NATURAL LANGUAGE PROCESSING"

Mengingat konten yang terdapat pada media sosial yang terus berkembang. Jumlah ujuran kebencian di media *online* juga meningkat dikarenakan skala *web* yang sangat besar, diperlukan metode yang secara otomatis dapat mendeteksi perkataan ujaran kebencian. Selama beberapa tahun terakhir, minat terhadap pendeteksian ujaran kebencian secara *online* dan khususnya otomatisasi tugas ini terus berkembang dengan seiringnya dengan dampak sosial dari fenomena tersebut. Penelitian ini memberikan gambaran singkat, komprehensif dan terstruktur dari deteksi ujaran kebencian otomatis dan menguraikan pendekatan yang ada secara sistematis. Dengan focus pada ekstrasi fitur secara khusus. Ini dilakukan untuk peneliti *Natural Language Processing* (NLP) yang baru pada bidang deteksi tersebut. Secara sistematis pada penelitian membandingkan fitur karakter n-gram dengan token n-gram untuk mendeteksi ujaran kebencian dan menemukan bahwa karakter n-gram terbukti lebih prediktif daripada token n-gram. Selain itu fitur yang berbasis kata dan karakter, pada deteksi tersebut juga dapat memanfaatkan fitur permukaan lainnya (Chen et al,.2012; Nobata et al., 2016). Seperti informasi tentang frekuensi penyebutan dan tanda baca pada URL, komentar, dan panjang token, kapitalisasi, kata-kata yang tidak dapat ditemukan dalam kamus bahasa. Serta jumlah karakter non-alfa numerik yang ada didalam token. Namun pendeteksian ujaran kebencian biasanya diterapkan

pada potongan kecil teks (bagian atau bahkan suatu kalimat), seseorang mungkin akan menghadapi masalah dikarenakan ketersebaran data. Banyak vector yang dapat digunakan sebagai fitur klasifikasi menggantikan fitur biner yang menunjukan keberadaan atau frekuensi kata tertentu. Karena dalam kalimat atau bagian deteksi ujaran kebencian diklasifikasikan daripada kata-kata individual, representasi vector dari kumpuln vector kata yang mewakili kata-kata dari teks yang akan diklasifikasikan akan dicari. Oleh karena itu dibutuhkan beberapa pendekatan dengan merujuk keterkaitan ujaran kebencian dan analisis sentimen dengan memasukan analis sentimen tersebut sebagai klasifikasi tambahan.

2.1.5 Literatur 5

(Nobata, et al., 2016) Penelitian ini berjudul "ABUSIVE LANGUAGE DETECTION IN ONLINE USER CONTENT"

Dalam penelitian ini, dilakukannya pengembangan metode berbasis pembelajaran mesin untuk dapat mendeteksi ujaran kebencian pada komentar pengguna online dari dua domain yang mengungguli pendekatan *deep learning*. Juga dijelaskan pengembangan tersebut diterapkan pada kumpulan pada komentar pengguna yang dianotasi untuk bahasa kasar. Setiap kali seseorang melakukan aktivitas *online,* baik di suatu forum papan pesan, melalui komentar, ataupun media sosial selalu ada risiko serius bahwa mungkin orang tersebut akan menjadi target celaan dan bahkan pelecahan. Kata-kata dan kalimat seperti menunjuk seseorang untuk melakukan hal-hal tidak wajar ataupun yang bertujuan negatif. Adalah hal yang tidak biasa secara *online* dan dapat berdampak besar pada kesopanan komunitas atau pengalaman para pengguna. Bahasa kasar sebenarnya mungkin sangat menjadi bagian suatu media sosial dan

hal tersebut tidak bias terlepas dengan begitu saja. Yang bisa menjadi tanda, bergunanya untuk metode otomatis dimana banyak kasus pada bahasa kasar atau bahkan lebih banyak ujaran kebencian. Inti pada penelitian ini menekankan bahwa bahasa kasar tidak terbatas hanya pada kalimat. Dalam beberap kasus, seseorang harus mempertimbangkan kalimat lain untuk memutuskan apakah teks tersebut kasar atau mengandung insiden perkataan yang mendorong kebencian. Dalam beberapa kasus lainnya beberapa pengguna memposting komentar sarkastik dengan suara yang sama dengan orang-orang yang membuat bahasa kasar.

2.1.6 Literatur 6

(Arifin, et al., 2014) Penelitian ini berjudul "EMOTION DETECTION OF TWEETS IN INDONESIAN LANGUAGE USING NON-NEGATIVE MATRIX FACTORIZATION"

Istilah indeks deteksi emosi pada *tweet* dalam bahasa Indonesia berupa emoji, *emoticon, hashtag* dan lainnya mungkin banyak digunakan oleh para pengguna. Deteksi emosi sendiri merupakan pemprosesan bahasa alami dan *text mining* dimana secara otomatis menunjukan emosi orang dari data tekstual seperti *tweet* atau cuitan. Beberapa penelitian mengenai pendeteksian emosi dengan menggunakan data berupa *tweet*, dikarenakan teks yang digunakan pendek serta sebagian besar bersifat informal sehingga dianggap sebagai masalah. Hingga saat ini, deteksi emosi teks pada penelitian ini dikelompokan kedalam enam kelas emosi dasar yaitu kemarahan, kegembiraan, kesedihan, ketakutan, rasa syukur, dan kejutan. Pada setiap *tweet* ini diberi label secara *manual* sesuai dengan kelas emosi sebelumnya. Secara umum para pengguna lebih banyak

memilih untuk menggunakan emoji untuk mengekspresikan emosinya, namun tidak cukup menutupi masalah pendeteksian emosi tersebut. Oleh karena itu pada fitur *emoticon* juga diterapkan. Ketidakjelasaan atau ambigu juga dapat terjadi seiringnya menggunakan emoji dan *emoticon* pada pelabelan data. Karena tidak semua emoji dan *emoticon* tersebut memiliki relevansi yang erat dalam emosi kelas. Dalam penelitian ini juga mengusulkan metode yang digunakan untuk deteksi emosi pada tweet berbahasa Indonesia dengan menggunakan *Negative Matrix Factorization* (NMF) untuk menentukan relevansi antar fitur secara tepat. Dengan mengumpulkan *tweet* pengguna yang berupa hashtag, emoji, emoticon, dan beberapa kata yang menunjukan emosi untuk penilaian relevansi. Lalu data-data tersebut akan melalui tiga proses setelah dikumpulkan, pengurangan data, dan penilaian relevansi emosi, Penentuan relevansi ini dan kelas emosi dilakukan secara manual oleh pakar berdasarkan topik. Dan juga dilakukannya pra-pemprosesan data yang merupakan fase dimana bentuk *tweet* informal dan mungkin terdapat karakteristik yang akan menjadi data pada akhirnya.

2.1.7 Literatur 7

(Abbas & Kareem, 2019) Penelitian ini berjudul "TEXT LANGUAGE INDENTIFICATION USING LETTERS (FREQUENCY, SELF-INFORMATION, AND ENTROPY) ANALYSIS FOR ENGLISH, FRENCH, AND GERMAN LANGUAGES"

Pada penelitian ini mencoba untuk dapat mendeteksi bahasa aslinya dan berhasil mendeteksi bahasa-bahasa tersebut. Setelah diterapkan pada suatu dokumen teks yang dipilih secara acak. Dengan meningkatkan komunikasi masih menjadi hal

yang sangat penting bagi komunitas dunia saat ini. Perkembangan pesat dalam komunikasi global telah menyebabkan kebutuhan akan sistem yang mampu mengklasifikasikan suatu bahasa pada dokumen dengan tepat. Oleh karena itu, penelitian ini mengenalkan deteksi bahasa asli serta mengukur tingkat keberhasilan dalam pendeteksian bahasa- bahasa tersebut pada dokumen teks. Dengan menggambarkan informasi sebagai ukuran kuantitatif dari suatu pertukaran komunikasi (Shannon, 1948). Teori ini menganalisis inforamsi berdasarkan teori probabilitas keberadaan tanda dalam konteksnya masing-masing. Bagian ini juga menyajikan beberapa prinsip dasar teori informasi selain itu definisi dan notasi probabilitas yang akan digunakan. Teori probabilitas merupakan salah satu bagian matematikan yang penting dalam perancangan sistem komunikasi digital. Meskipin frekuensi relative suatu peristiwa merupakan gagasan terapan, hal itu mengacu pada waktu peristiwa yang akan benar-benar terjadi. Pada bidang frekuensi huruf, karena pola huruf dalam bahasa tidak terjadi secara acak maka frekuensi huruf dapat ditandai dalam beberapa aturan dan pola frekuensi huruf dapat membantu membedakan teks acak dari teks bahasa alami. Frekuensi huruf dalam teks sering dipertimbangkan untuk pemerikasaan kriptografi dan frekuensi yang disetujui. Tidak ada distribusi frekuensi huruf yang tepat yang mendasari bahasa tertentu karena dokumen-dokumen tersebut ditulis sampai dengan batas tertentu secara berbeda. Analisis lainnya menunjukan bahwa frekuensi huruf seperti frekuensi kata, memiliki kecenderungan berbeda menurut subjek dan penulis. Sedangkan dalam teori informasi dapat dijelaskan sebagai kuantitas informasi dengan pengetahuan tentang hasil dari suatu peristiwa. Secara umum, kuantitas informasi ini dalam

peristiwa probabilistic hanya dapat bergantung pada probabilitas peristiwa tersebut.

2.1.8 Literatur 8

(Chavez, et al., 2015)Penelitian ini berjudul" FAST ACCURATE LANGUAGE DETECTION IN SHORT TEXT USING CONTEXTUAL ENTROPY"

Media sosial menjadi tempat mendapatkan informasi. Yang menjadi bagian penting yang menjadi *corpus* berupa teks pendek. Ide utama penelitian ini adalah menghitung entropi suatu dokumen dalam konteks dan meletakannya pada kategori dimana entropi maksimal. Untuk *Language Indentification* (LI) dengan konteksnya adalah bahasa dan dilakukan hanya dengan mengevaluasi entropi tingkat tinggi dari teks-teks tersebut. Lalu hasil yang ditunjukan berupa bahasa teks, dalam kasus teks pendek dapat secara akurat mencocokan pendekatan yang mampu menginformasikan kedalam sebuah literatur. Mengingat beberapa data pada hasil sebelumnya serta *data training* untuk setiap teks terdapat label dimana menunjukan bahasa pada teks yang ditulis. Tujuannya adalah untuk dapat mempelajari model sedemikian rupa sehingga dengan beberapa teks yang sebelumnya tidak terlihat dapat teridentifikasi sekakurat mungkin dengan suatu bahasa pada teks yang ditulis. Beberapa kasus untuk mengklasifikasikan *Language Indentification* yaitu, sebuah teks yang ditulis sebagian dalam satu bahasa dan sebagian lagi dalam beberapa bahasa lain dan seseorang ingin mendapatkan kedua label pada hasil yang akan diterima. Sedangkan pada pembahasan lainya tedapat entropi. Dimana Entropi sendiri adalah statistic yang bergantung pada objek itu sendiri (teks) dan konteksnya (distribusi kosa kata).

2.1.9 Literatur 9

(Pratiwi, et al., 2017) Penelitian ini berjudul "STUDY OF HOAX NEWS DETECTION USING NAÏVE BAYES CLASSIFIER IN INDONESIA LANGUAGE"

Hingga saat ini media internet telah banyak diketahui sebagai salah satu sumber informasi termasuk salah satunya adalah artikel berita *online*. Pada umum nya orang-orang akan mencari suatu berita dengan internet. Dan artikel-artikel tersebut tersebar dan banyak pada setiap *website*. Artikel-artikel tersebut juga dapat berupa asli ataupun palsu. Dimana berita artikel palsu itu disebut juga sebagai berita *hoax*. Berita hoax hanya akan akan berniat menipu para pengguna yang membacanya, ataupun akan meprovokasi seseorang untuk hal buruk lainnya. Pada penelitian ini bertujuan untuk membangun suatu deteksi otomatis mengenai *hoax* dengan mengindetifikasi berita-berita pada bahasa Indonesia. Pada masalah tersebut, tujuan pembelajaran ditunjukan untuk membuat suatu *dataset* yang valid dan berita *hoax* pada suatu artikel dengan menggunakan klasifikasi dengan algoritma *machine learning* yaitu *naïve bayes* yang nantinya akan sangat berguna untuk pengembangan kedepannya. Pada penelitian ini menggunakan *naïve bayes* yang dimana sangat popular untuk digunakan pada machine learning sebagai klasifikasi, sederhana, tingginya pempromresan dan juga akurasi yang baik.

2.1.10 Literatur 10

(Flores, et al., 2011) Penelitian ini berjudul"TOWARDS THE DETECTION OF CROSS-LANGUAGE SOURCE CODE REUSE"

Pada era digital ini, sejumlah besar informasi yang tersedia yang menyebabkan materi dari orang lain diekspos untuk dapat digunakan kembali. Oleh karena itu, adanya suatu keinginan untuk dapat mengidentifikasi apakah suatu karya telah digunakan kembali. Adapun beberapa dokumen dalam bahasa alami dan sejumlah kode pada sumber di internet. Dengan memfasilitasi penggunan semua ataupun sebagian dari program yang telah dibuat sebelumnya. Penelitian ini tidak mereferensikan ke karya asli yang disertakan atau plagiarism. Lalu banyaknya juga model dalam penggunaan kembali sumber tersebut. Contoh beberapa diantaranya adalah tidak adanya perubahan pada *source code*, menghadirkan perubahan pada komentar dan indentasi. Evaluasi ini akan dilakukan untuk mendeteksi pada *cross-language* yang digunakan kembali khususnya pada *source code*.

2.2 Tinjauan perbedaan penelitian ini dengan penelitian sebelumnya

Berdasarkan beberapa literatur pada penelitian sebelumnya, diantaranya menggunakan deteksi emosi dengan pemprosesan bahasa alami dan *text mining* dengan cara mendapatkan teks secara otomatis yang dapat diolah menjadi *data* tekstual. Dengan pengelompokan kembali pada jenis-jenis emosi yang dapat diukur namun beberapa fitur pada deteksi tersebut tidak dapat diterapkan seperti fitur *emoticon* yang digunakan oleh pengguna. Dengan fitur tersebut beberapa hal dapat menjadi ambigu seiiringnya menggunakan emoji dan *emoticon* karena

secara keseluruhan emoji dan *emoticon* tidak memiliki keterikatan yang erat pada emosi dan tidak dapat menjadi nilai acuan yang dapat di definisikan. Sedangkan untuk penelitian ini dibuat dengan metode *naïve bayes*, dimana *data* yang dikumpulkan berupa korpus yang dapat diolah menjadi *dataset*. Yang akan digunakan untuk melatih membangun sebuah model untuk dapat mendeteksi bahasa berdasarkan korpus yang digunakan. Tinggi atau rendahnya kualitas dan jumlah korpus juga menjadi salah satu faktor tingkat keberhasilan dan hasil kualitas dalam mendeteksi bahasa tersebut. Melalui beberapa pendekatan pada literatur sebelumnya, masih terdapat pengambilan data secara manual, acak, dan otomatis. Selain *text mining*, literatur lainnya juga menggunakan *deep learning*, pembuatan vektor kata pada pemprosesan bahasa alami, *mining bilingual text*, *random forest decision tree classifier*, dan sebagainya dalam pemprosesan serta pengolahan kata.

2.3 Landasan Teori

Berdasarkan penelitian yang peneliti lakukan maka peneliti menyusun landasan teori yang berkaitan dengan penelitian yaitu sebagai berikut:

2.3.1 Naïve Bayes

Naïve Bayes adalah algoritma yang mengasumsikan indepedensi diantara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan kata dan informasi konteks dalam kalimat atau dokumen secara umum. Selain itu metode ini memperhitungkan jumlah kemunculan kata dalam dokumen (Destuardi & Surya, 2009). Asumsi tersebut pada keindenpendanan atribut pada data jarang terjadi, walaupun asumsi keindependenan atribut dilanggar, klasisfikasi ini memiliki performa pengklasifikasian yang cukup tinggi. Klasifikasi *Naïve Bayes* termasuk kedalam algoritma pembelajaran *Bayes*. Algoritma ini dibangun dari *data train* untuk menentukan probabilitas untuk menentukan bagaiman setiap probabilitas pada setiap kategori yang terdapat pada ciri dokumen yang sedang diuji. *Data train* yang digunakan untuk melatih sistem memahami pengkategorian suatu data (Maryamah,2016).

2.3.2 Natural Language Processing (NLP)

Natural Language Processing atau lebih dikenal dengan Pemprosesan Bahasa Alami (PBA) adalah suatu kemampuan pada komputer yang dapat memproses atau mengolah berupa bahasa, baik dalam bentuk lisan ataupun tulisan yang digunakan oleh manusia dalam percakapan sehari-hari. Dalam proses komputsinya, suatu bahasa direpresentasikan kedalam serangkaian symbol dengan memenuhi aturan tertentu. Atau dapat dikatakan bahwa, pemprosesan bahasa tersebut dibuat untuk komputer agar dapat mengerti perintah-perintah yang digunakan ataupun yang ditulis dalam standar

bahasa yang digunakan oleh manusia. Pada NLP juga sering ditemukan beberapa hal yang menyulitkan (Arman, 2004), yaitu seperti pada masalah *ambiguity* atau makna ganda dan juga jumlah kosa kata (vocabulary) yang besar dengan seiringnya perkembangan waktu ke waktu.

2.3.3 Python

Python merupakan salah satu bahasa pemrograman yang populer di dunia kerja. Selain itu, di ranah akademik pun banyak akademisi yang menggunakan python untuk menyelesaikan penelitiannya dibidang komputasi sains, robotika, *data science*, dan bidang lainnya. Python secara *default* telah terpasang di beberapa sistem operasi berbasis Linux, seperti Ubuntu, Linux Mint, dan Fedora. Untuk sistem operasi lain, sudah tersedia *installer* yang disediakan untuk sistem operasi tersebut.

Bahasa pemrograman Python merupakan bahasa yang sangat singkat dalam koding atau *script* sehingga banyak pengembang yang menggunakannya. Banyaknya juga paket-paket yang dibuat untuk mendukung Python dalam pembuatan aplikasi, sehingga aplikasi ini juga sangat baik untuk digunakan dengan aplikasia lain. Contoh beberapa diantaranya:

1. Scipy dan Scikit

   Library atau paket *python* untuk membuat aplikasi *machine learning* dan kecerdasan buatan (*artificial Intelligence*).
2. OpenCV Python

   Library atau paket *python* untuk membuat aplikasi computer *vision.*
3. BioPython

   Library atau paket *python* untuk menganalisa DNA atau *genome* makhluk hidup.

4. Tornado

   Library atau paket *python* untuk membuat aplikasi *web*, *websocket*, dan *asynchronous programming*.

5. Matplotlib

   Library atau paket *python* untuk membuat grafik untuk keperluan saintifik.

6. Celery

   Library atau paket *python* untuk membuat *asynchronous task*.

7. TensorFlow

   Library atau paket *python* untuk membuat aplikasi yang ditenagai oleh *deep learning*.

8. Django

   Library atau paket untuk membuat *web* (*web framework*).

Dan masih banyak paket (pustaka) / *library* yang dibuat menggunakan *python*. *Python* juga memiliki sebuah *package manager* unggul dan popular yang dinamakan PIP. Dengan menggunakan PIP tersebut, para pengguna *python* dapat memulai pemasangan atau menghapus pustaka *python* yang akan atau tidak lagi digunakan(Supardi,2017).

2.3.4 N-Gram

N-Gram adalah suatu urutan n unit yang pada umumnya berupa karakter tunggal atau string yang dipisahkan oleh spasi. . N-Gram adalah potongan N-karakter yang diambilkan dari suatu string. Blank ditambahkan pada awal dan akhir suatu string untuk mengetahui batas awal dan akhir suatu string. Sebagai contoh pada suatu string

“KATA” dengan mengganti blank dengan “_” pada awal dan akhir maka akan didapat N-Gram sebagai berikut:

Unigram : K,A,T,A

Bigram : _K, KA, AT, TA, dan A_

Trigram : _KA, KAT, ATA, TA_, dan A_ _

Dapat disimpulkan bahwa untuk string berukuran n akan dimiliki pada n unigram dan n+1 bigram, n+1 trigram dan seterusnya. Penggunaan N-Gram untuk *matching* kata memiliki keuntungan sehingga dapat diterapkan pada *recovery* pada input karakter ASCII yang terkena *noise*, interpretasi kode pos, *information retrieval* dan berbagai aplikasi dalam pemrosesan bahasa alami. Dengan model bahasa bigram memproses bahasa dengan memecah kalimat menjadi beberapa bagian dimana tiap bagian terdiri dari dua kata terurut dari kalimat. Jika suatu kalimat dinyatakan dalam notasi $\{w_1, w_2, w_3..., w_{i-1}, w_i\}$ dan *i* menyatakan banyak kata dalam kalimat, maka nilai peluang bigram dapat dihitung melalui persamaan berikut,

$$P(w_i|w_{i-1}) = \frac{N(w_{i-1}, w_i)}{N(w_{i-1})}$$

Notasi $P(w_i|w_{i-1})$ menyatakan nilai peluang bigram kata ke-i, didahului oleh kata ke (i-1). Notasi $N(w_{i-1},w_i)$ menyatakan banyaknya kemunculan pasangan kata ke-i didahului oleh kata ke (i-1) pada korpus. Notasi $N(_{wi-1})$ menyatakan banyaknya kemunculan kata ke (i-1) pada korpus. Sebuah kalimat dapat dihitung nilai peluangnya menggunakan model statistik bigram dengan menerapkan aturan Chain yang dijabarkan dalam Persamaan berikut.

$$P(W_1^n) \approx \Pi_{i=1}^{n} P(w_i/w_{i-1})$$

Notasi $P(W_1^n)$ Menyatakan nilai peluang kalimat yang dihitung menggunakan model statistik bigram. Notasi n menyatakan banyaknya pasangan bigram kata dalam sebuah kalimat. Contoh penerapan persamaan adalah sebagai berikut.

W = Saya suka pelajaran matematika

Maka peluang bigram dari kalimat tersebut dapat dihitung dengan rumus.

P(W) = P(suka | saya) * P(pelajaran | suka) * P(matematika | pelajaran)

2.3.5 PEM (Perfomance Evaluation Measure)

Pengujian model akan dilakukan dengan menghitung 3 pendekatan PEM yaitu *accuracy*, *precision*, dan *recall*. Dan lebih dikenal dengan *Confusion Matrix*, dengan menggunakan *matrix* 2x2 yang merepresentasikan hasil dari klasikasi. Maka dapat dihitung kedalam tiga pendekatan tersebut pada persamaan berikut. Yang dapat dilihat pada Gambar 2.1 Pendekatan PEM dan Confusion Matrix

a. Accuracy

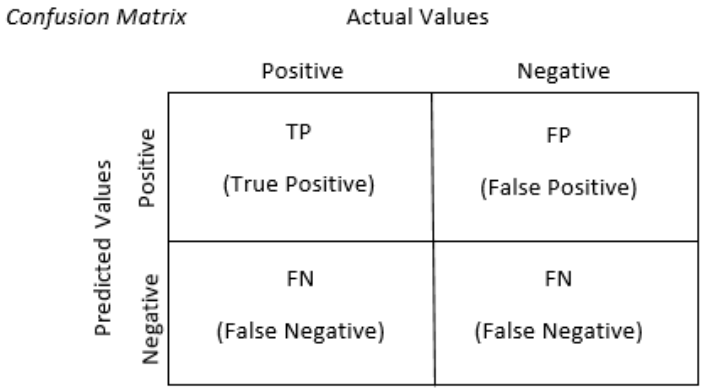
Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

b. Precision

Precision =$\frac{TP}{TP+FP}$

c. Recall

Recall =$\frac{TP}{TP+FN}$



Gambar 2.1 Pendekatan PEM dan Confusion Matrix

Dengan sebuah *classifier* dan sebuah instansi, yang akan mempunyai empat kemungkinan nilai *output*. Jika instansi tersebut bernilai positif lalu diklasifikasikan kedalam kelompok positif maka akan dihitung sebagai *true positive*, jika diklasifikasikan kedalam kelompok negatif maka akan dihitung sebagai *false negative*. Sedangkan untuk instansi yang bernilai negatif lalu di klasifikasikan kedalam kelompok negatif maka akan dihitung sebagai *true* negative, jika diklasifikasikan kedalam kelompok positif maka akan dihitung sebagai false *positive*.

Accuracy adalah sebuah nilai keakuratan antara informasi yang diolah dan akan ditinjau kembali untuk melihat perbandingan pada setiap bagian pada informasi tersebut.

Precision adalah sebuah kecocokan atau keserupaan (antara permintaan informasi dengan jawaban pada permintaan tersebut). *Precision* adalah jumlah

kelompok dokumen yang relevan dari total jumlah dokumen yang akan ditemukan oleh sistem.

Recall dapat diartikan sebagai proporsi pada jumlah dokumen – dokumen yang dapat ditemukan kembali melalui sebuah proses pencarian pada sistem *Information Retrieval*. Dengan kemampuan suatu sistem pada temu balik dalam menemukan dokumen yang relevan.