

Natural language processing (NLP)

Disampaikan Oleh : Dodo Zaenal Abidin



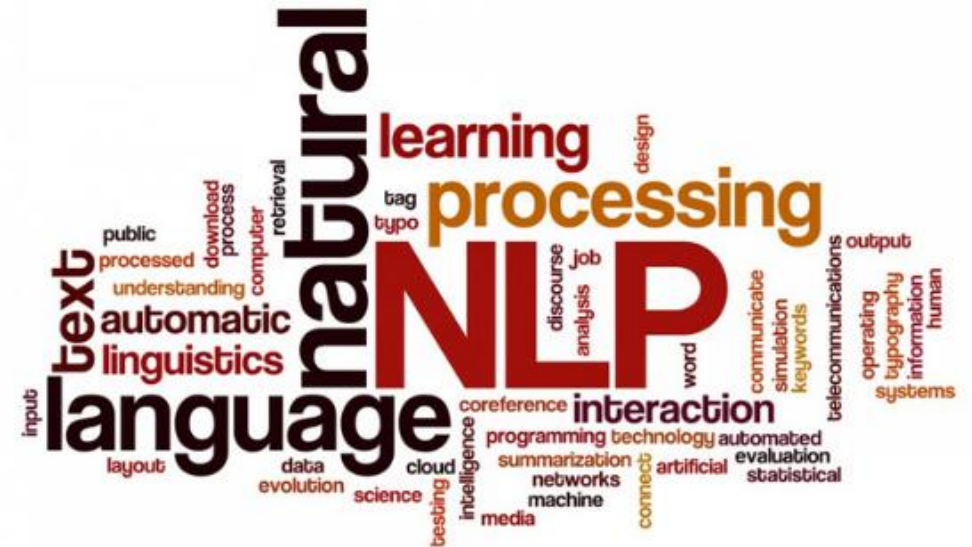
https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

Apa Itu *natural language processing*?

Natural language processing (NLP) adalah sub-bidang ilmu komputer, teknik informasi, dan *artificial intelligence* yang berkaitan dengan interaksi antara komputer dan bahasa manusia (asli). Ini tidak lain adalah bagaimana memprogram komputer untuk memproses dan menganalisis data bahasa alami dalam jumlah besar.

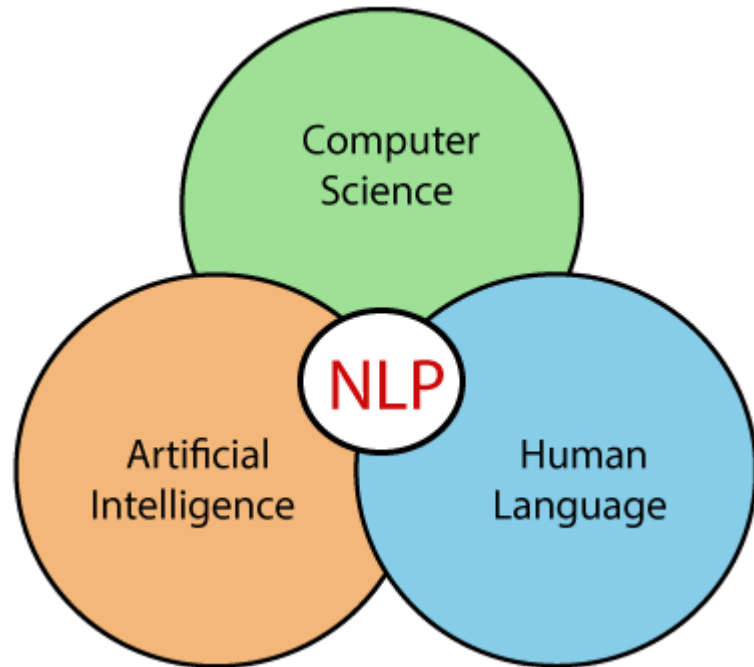
NLP adalah kemampuan perangkat lunak komputer untuk memahami bahasa manusia saat diucapkan.

Natural language processing (NLP) mengacu pada cabang ilmu komputer — dan lebih khusus lagi, cabang *artificial intelligence* atau AI — yang berkaitan dengan memberi komputer kemampuan untuk memahami teks dan kata-kata yang diucapkan dengan cara yang sama seperti manusia. (<https://www.ibm.com/id-en/topics/natural-language-processing>)



<https://dzone.com/articles/natural-language-processing-1>

Hubungannya dengan CS/AI/HL

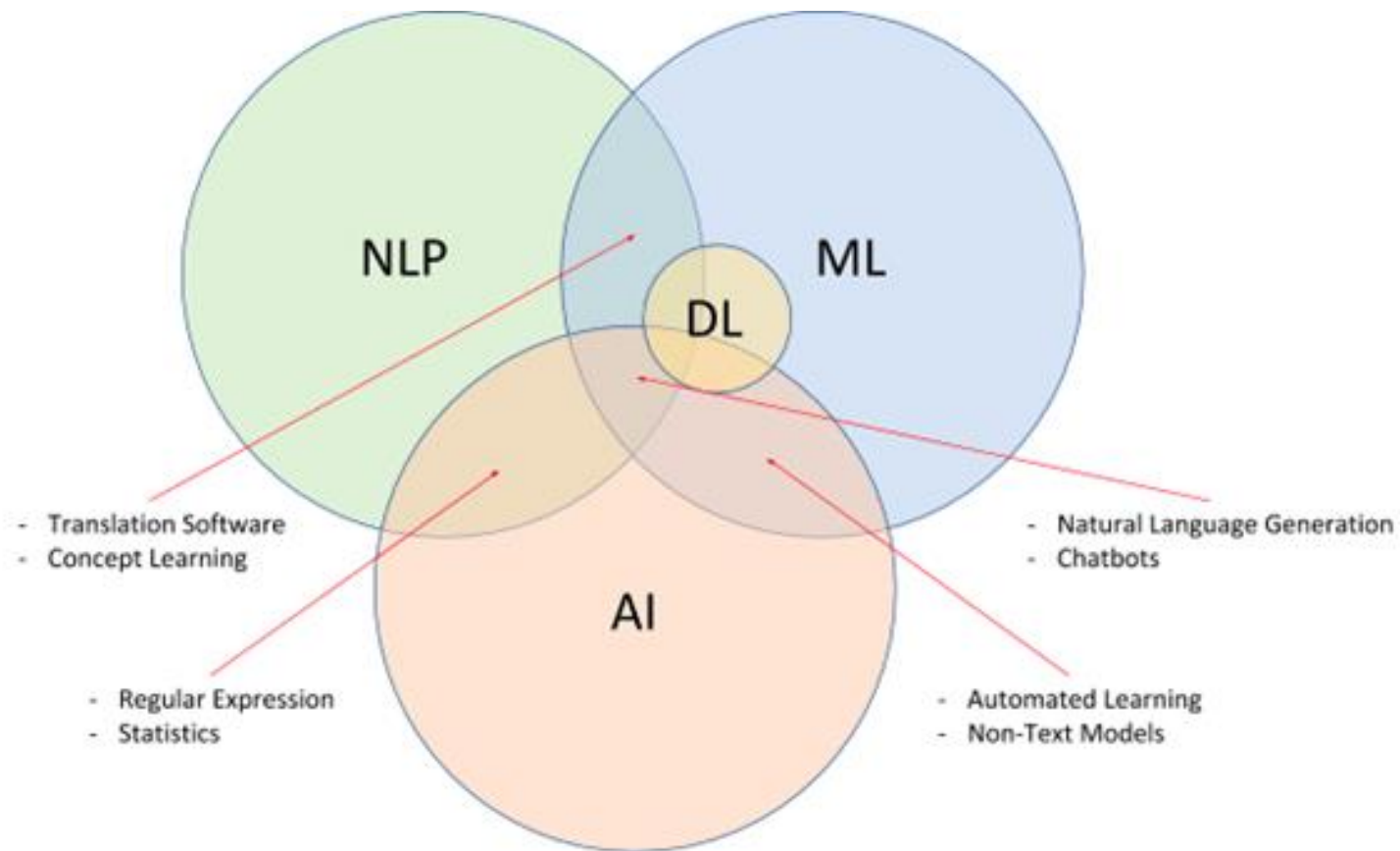


<https://www.javatpoint.com/nlp>

Natural Language Processing, yang merupakan bagian dari Ilmu Komputer (*Computer science*), Bahasa Manusia (*Human language*), dan Kecerdasan Buatan (*Artificial Intelligence*) .

Ini adalah teknologi yang digunakan oleh mesin untuk memahami, menganalisis, memanipulasi, dan menginterpretasikan bahasa manusia. Ini membantu pengembang untuk mengatur pengetahuan untuk melakukan tugas-tugas seperti *translation*, *automatic summarization*, *Named Entity Recognition (NER)*, *speech recognition*, *relationship extraction*, dan *topic segmentation*.

Hubungannya dengan NLP/AI/ML/DL



<https://thinkml.ai/what-are-the-natural-language-processing-challenges-and-how-to-fix-them/>

Sejarah NLP



A BRIEF HISTORY OF NATURAL LANGUAGE

Natural Language Processing (NLP) is a type of artificial intelligence which aids computers to understand human language and communicate in human-like ways. This infographic pinpoints key historical events which have supported NLP advancements.

Machine translation used in hopes to **break codes in WW2**, translating Russian into English. Results are unsuccessful.

Noam Chomsky releases the **Syntactic Structures** which advances linguistic studies with a universal grammar rule.

1940S

1957

1600S:

Philosophers **Leibniz** and **Descartes** propose **theoretical codes** in relation to **language**.

1930S

Patents are submitted for 'translating machines'. **George Artsrouni** applies to build an **automatic bilingual dictionary**. **Peter Troyanskii** proposes another dictionary that processes variations in grammar across languages.

1950

Alan Turing publishes 'Computing Machinery and Intelligence' which outlines concept of **Turing test**.

SHRDLU, an **early NLP** program, developed by **Terry Winograd** at **MIT** which allows computers and people to converse but with restrictions.

1968-1970

The first **statistical machine translation systems** are developed. Strict and complex hand-written rules are swapped for **newly-developed algorithms** which increase a computer's understanding.

1980S

IBM create **AI software**, **Watson** which goes on to win competition against best human contestants in 2011.

2006

Rising **adoption rates of AI-powered bots** for customer-facing roles. NLP will continue to develop so communication with computers will be as effortless as human interactions.

2020 +

1966

ELIZA, a computer psychotherapist and **first bot**, is created by **Joseph Weizenbaum**.

1970-1980

Roger Schank introduces conceptual dependency theory for NLP. **William A. Woods** releases the **augmented transition network** to show natural language inputs. A wealth of bots are written including **PARRY**.

1990-2000S

Programmers develop **models** to increase the capabilities of computers using NLP.

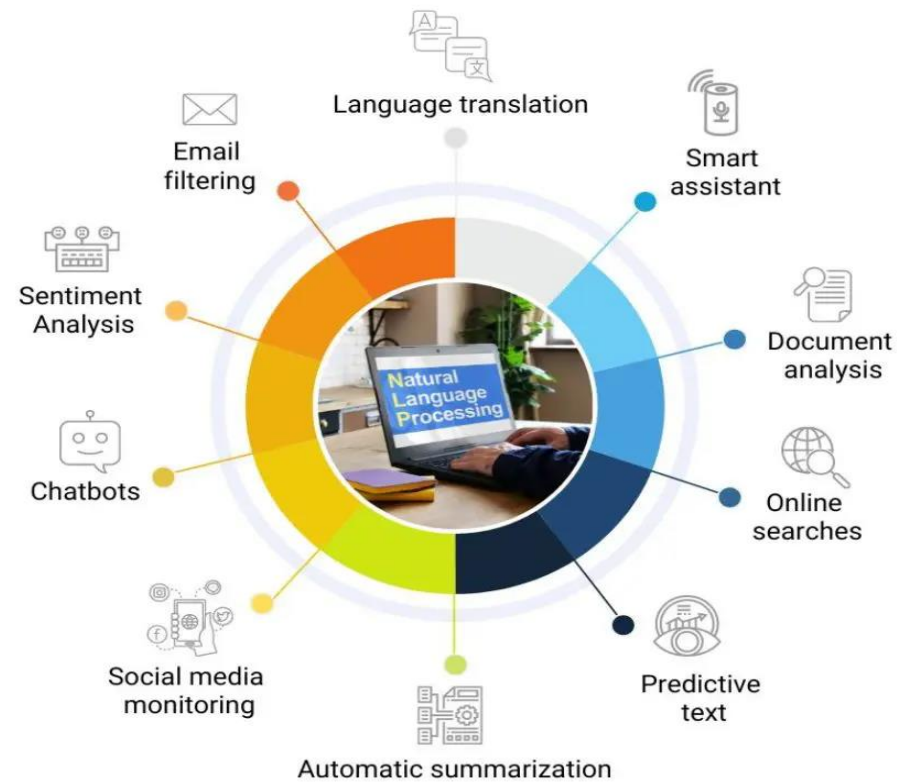
2010-2020

People introduce technologies that utilise **NLP into their homes**, such as mobile assistant **Siri** (2011) and Amazon assistant, **Alexa** (2014). 2017 marks the **rise in chatbot integration** into business operations.

<https://www.hostcomm.co.uk/blog/2019/evolution-natural-language-infographic>

Penerapan *Natural Language Processing*

Applications of Natural Language Processing



<https://datasciencedojo.com/blog/natural-language-processing-applications/>

Penerapan *Natural Language Processing*

1. *Email filtering*

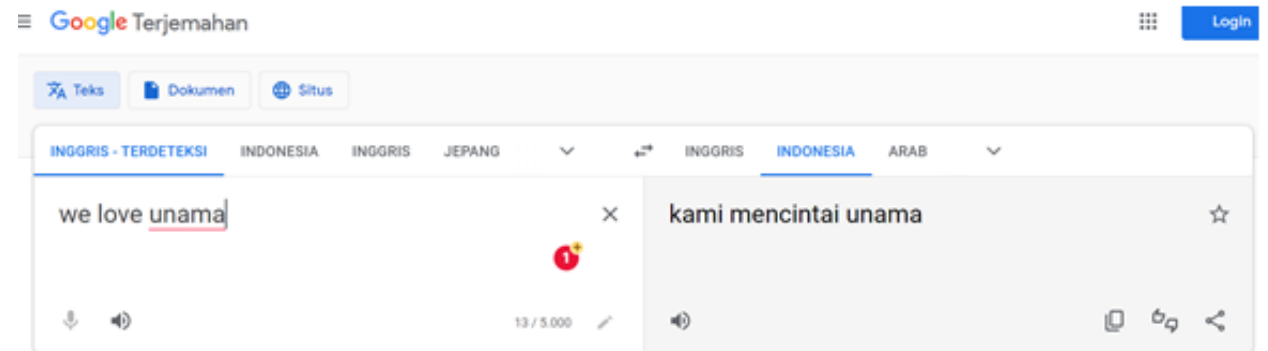
Email adalah bagian dari kehidupan sehari-hari. Apakah itu terkait dengan pekerjaan atau studi. Di sini **NLP** bekerja untuk mengidentifikasi dan memfilter email yang masuk ke dalam "penting" atau "spam" dan menempatkannya ke dalam peruntukannya masing-masing.



<https://www.mvttwireless.com/why-are-my-emails-going-into-spam/>

2. *Language translation*

Ada banyak bahasa di dunia ini, kita perlu berkomunikasi dengan orang lain yang berbicara bahasa yang mungkin asing bagi kita. **NLP** membantu untuk menerjemahkan bahasa dengan semua sentimennya.



Penerapan *Natural Language Processing*

3. Smart assistants

Teknologi yang cukup canggih untuk memiliki asisten cerdas (*smart assistants*), seperti Siri, Alexa, dan Cortana. Kita dapat berbicara dengan mereka seperti kita berbicara dengan manusia normal, dan mereka bahkan menanggapi kita dengan cara yang sama.

NLP membantu sistem komputer memahami bahasa kita dengan memecahnya menjadi bagian-bagian ucapan (*parts of speech*), batang akar (root stem), dan fitur linguistik lainnya. Itu tidak hanya membantu mereka memahami bahasa tetapi juga dalam memproses makna dan sentimennya dan menjawab kembali dengan cara yang sama seperti manusia.

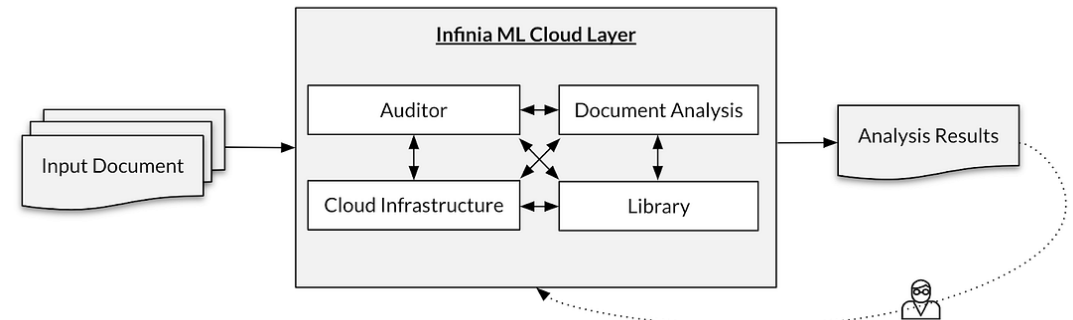


<https://grbizm.com/market-power-of-siri-alexa-google-a-concern-eu-regulators-say/>

Penerapan *Natural Language Processing*

4. Document analysis

Salah satu aplikasi **NLP** lainnya adalah analisis dokumen. Perusahaan, perguruan tinggi, sekolah, DLL selalu dipenuhi dengan data, yang perlu disortir dengan baik, dipelihara, dan dicari. Semua ini bisa dilakukan dengan menggunakan **NLP**. tidak hanya mencari kata kunci tetapi juga mengkategorikannya sesuai dengan instruksi dan menyelamatkan kita dari pekerjaan panjang dan sibuk mencari informasi satu orang dari tumpukan file. Aplikasi **NLP** juga membantu penggunaannya untuk menginformasikan pengambilan keputusan tentang klaim dan manajemen risiko.



<https://huxiangqian.medium.com/document-analysis-is-more-than-processing-text-b33eac15c5fb>

Penerapan *Natural Language Processing*

5. *Online searches*

Di dunia yang penuh dengan tantangan dan teka-teki ini, kita harus terus mencari jalan dengan mendapatkan informasi yang dibutuhkan dari sumber yang tersedia. Salah satu **sumber informasi yang paling luas** adalah **internet**. Tapi apakah Anda pernah berpikir tentang bagaimana Anda mendapatkan hasil tersebut bahkan ketika Anda tidak mengetahui kata kunci yang tepat yang Anda butuhkan untuk mencari informasi yang dibutuhkan?

NLP membantu mesin pencari (*search engine*) memahami apa yang diminta dari mereka dengan memahami arti literal dari kata-kata dan maksud dibalik penulisan kata itu, sehingga memberi kita hasil



Google

office supplies

Google Search

I'm Feeling Lucky

<https://edu.gcfglobal.org/en/search-better-2018/introduction-to-searching-online/1/>

Penerapan *Natural Language Processing*

6. *Predictive text*

Aplikasi serupa untuk **pencarian online** adalah **teks prediktif**. **teks prediktif** adalah sesuatu yang kita gunakan setiap kali mengetik apa pun di *smartphone*. Setiap kali mengetik beberapa huruf di layar, *keyboard* akan menyarankan apa kata berikutnya.

Semua ini dilakukan dengan menggunakan **NLP** dengan menjadikan *smartphone* kita cukup cerdas untuk menyarankan kata-kata dan belajar dari kebiasaan mengirim SMS, chat WA, dll.



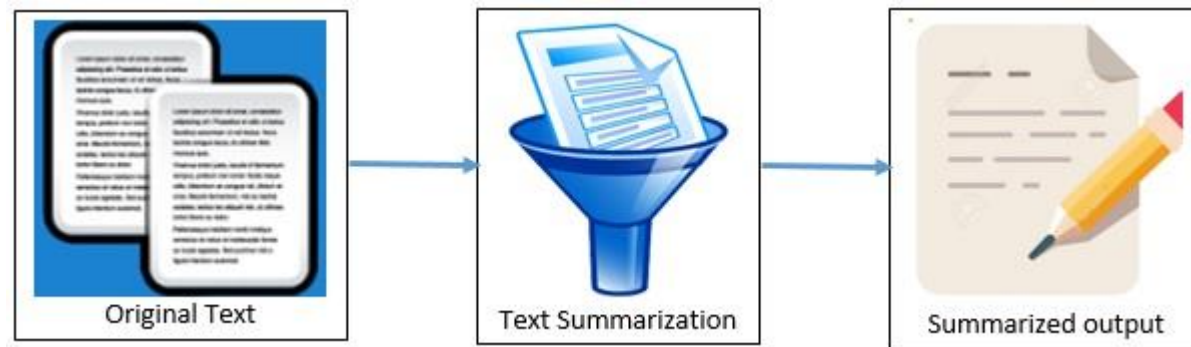
<https://www.samsung.com/uk/support/mobile-devices>

Penerapan *Natural Language Processing*

7. *Automatic summarization*

Dengan meningkatnya penemuan dan inovasi, data juga meningkat. **Peningkatan data** ini juga telah memperluas cakupan **pemrosesan data**. Namun, **pemrosesan data manual memakan waktu** dan rentan terhadap kesalahan.

NLP memiliki solusi untuk itu juga, tidak hanya dapat merangkum makna informasi, tetapi juga dapat memahami makna emosional yang tersembunyi di dalam informasi tersebut. Dengan demikian, **NLP** mampu membuat proses ringkasan cepat dan sempurna.



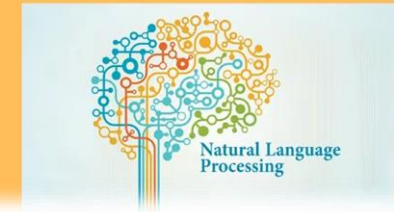
<https://www.analyticsvidhya.com/blog/2020/12/tired-of-reading-long-articles-text-summarization-will-make-your-task-easier/>

Penerapan *Natural Language Processing*

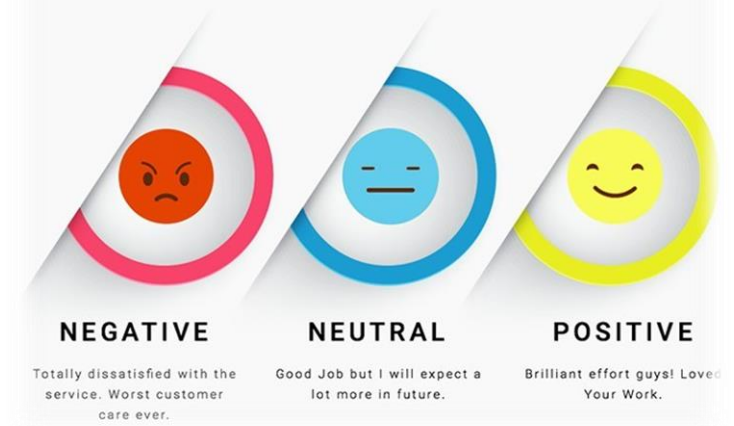
8. *Sentiment analysis*

Percakapan sehari-hari, **konten** dan **komentar** yang **diposting**, buku, restoran, dan ulasan produk, karenanya hampir semua **percakapan** dan **teks** penuh dengan **emosi**. Memahami **emosi** ini sama pentingnya dengan memahami **makna kata** demi kata. Kita sebagai manusia dapat menafsirkan **sentimen emosional** dalam **tulisan** dan **percakapan**, tetapi dengan bantuan **NLP**, sistem komputer juga dapat memahami **sentimen teks** beserta **makna literalnya**.

Sentiment Analysis using Python



SENTIMENT ANALYSIS



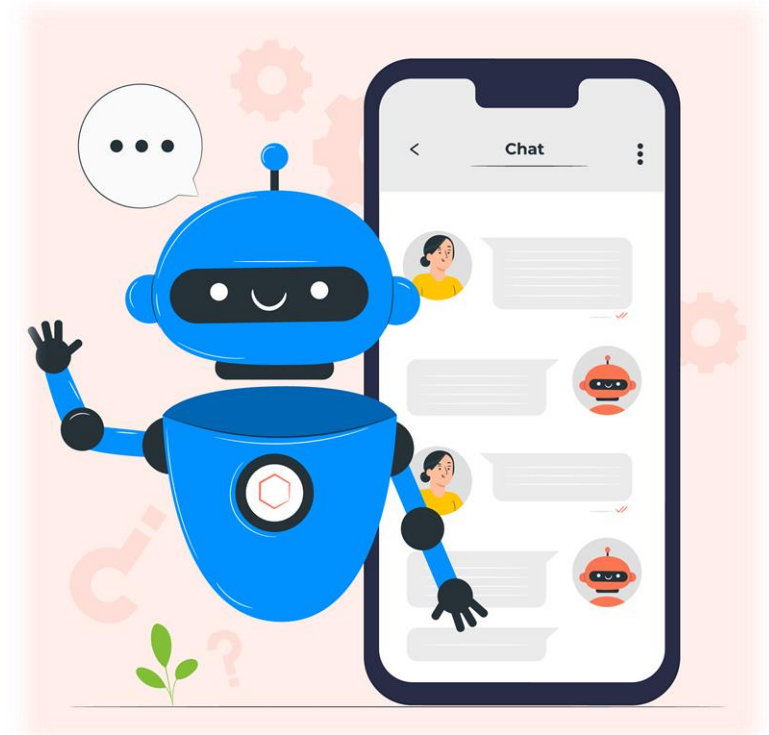
<https://www.askpython.com/python/sentiment-analysis-using-python>

Penerapan *Natural Language Processing*

9. *Chatbots*

Dengan peningkatan teknologi, semuanya telah didigitalkan, mulai dari belajar hingga berbelanja, memesan tiket, dan layanan pelanggan. Alih-alih menunggu lama untuk mendapatkan jawaban singkat dan instan, **chatbot** membalas secara instan dan akurat. **NLP** memberikan kemampuan percakapan **chatbots** ini, yang membantu mereka **merespons** dengan tepat kebutuhan pelanggan alih-alih hanya balasan sederhana.

Chatbots juga membantu di tempat-tempat di mana tenaga manusia kurang atau tidak tersedia sepanjang waktu. **Chatbots** yang beroperasi di **NLP** juga memiliki **kecerdasan emosional**, yang membantu mereka memahami **sentimen emosional** pelanggan dan meresponsnya secara efektif.



<https://appinventiv.com/blog/building-chatbot-using-nlp/>

Penerapan *Natural Language Processing*

10. *Social media monitoring*

Saat ini, setiap orang memiliki akun **media sosial** tempat mereka berbagi pemikiran, suka, tidak suka, pengalaman, dll., Yang menceritakan banyak tentang individu. Kami tidak hanya menemukan **informasi** tentang **individu** tetapi juga tentang **produk** dan **layanan**. Perusahaan terkait dapat mengolah data ini untuk mendapatkan informasi tentang **produk** dan **layanan** mereka untuk meningkatkan atau mengubahnya. NLP berperan di sini. Ini memungkinkan **sistem komputer** untuk **memahami** data media sosial yang tidak terstruktur, menganalisisnya, dan menghasilkan hasil yang diperlukan dalam bentuk yang berharga bagi perusahaan.



Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

□ Kata dan frasa kontekstual dan homonim

Kata dan frasa yang sama dapat memiliki arti yang berbeda sesuai dengan konteks kalimat dan banyak kata – terutama dalam bahasa Inggris – memiliki pengucapan yang sama persis tetapi arti yang sama sekali berbeda.

Misalnya: *I ran to the store because we ran out of milk.*
Can I run something past you real quick?
The house is looking really run down.

Homonim – dua kata atau lebih yang dilafalkan sama tetapi memiliki definisi yang berbeda – dapat menimbulkan masalah untuk menjawab pertanyaan dan aplikasi ucapan-ke-teks (*speech-to-text*) karena tidak ditulis dalam bentuk teks. misalnya, Penggunaan kata *there* dan *their* , bahkan merupakan masalah umum bagi manusia.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

□ Kata Sinonim

Sinonim dapat menyebabkan masalah yang mirip dengan pemahaman kontekstual karena menggunakan banyak kata berbeda untuk mengungkapkan ide yang sama. Selain itu, beberapa dari kata-kata ini mungkin menyampaikan arti yang persis sama, sementara beberapa mungkin memiliki tingkat kerumitan (*small, little, tiny, minute*) dan orang yang berbeda menggunakan **sinonim** untuk menunjukkan arti yang sedikit berbeda dalam kosa kata pribadi mereka.

Dalam membangun sistem **NLP**, penting untuk menyertakan semua kemungkinan arti kata dan semua **sinonim** yang mungkin. Model analisis teks terkadang masih membuat kesalahan, tetapi semakin relevan data pelatihan yang mereka terima, semakin baik mereka dapat memahami **sinonim**.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

□ Kata **Ironi** dan **sarkasme**

Ironi dan **sarkasme** menghadirkan masalah bagi model pembelajaran mesin karena mereka umumnya menggunakan kata dan frasa yang, menurut definisi, mungkin positif atau negatif, tetapi sebenarnya berkonotasi sebaliknya.



Model dapat dilatih dengan isyarat tertentu yang sering menyertai frasa ironis atau sarkastik, seperti "**yeah right**", "**whatever**", dll., dan penyematan kata (di mana kata-kata yang memiliki makna yang sama memiliki representasi yang serupa), tetapi tetap rumit proses.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

□ Kata Ambiguity

Ambiguitas dalam NLP mengacu pada kalimat dan frasa yang berpotensi memiliki dua atau lebih kemungkinan interpretasi.

- **Lexical ambiguity** : kata yang dapat digunakan sebagai kata kerja (verb), kata benda (noun), atau kata sifat (adjective).
- **Semantic ambiguity** : interpretasi kalimat dalam konteks. Misalnya: *"I saw the boy on the beach with my binoculars. This could mean that I saw a boy through my binoculars or the boy had my binoculars with him"*
- **Syntactic ambiguity** : Pada kalimat di atas, inilah yang menimbulkan kerancuan makna. Ungkapan dengan teropong saya (*with my binoculars*) dapat mengubah kata kerja, *"saw"*, atau kata benda, *"boy"*.

“Penandaan POS (*part of speech*) adalah salah satu solusi NLP yang dapat membantu menyelesaikan masalah”.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Kesalahan dalam teks dan ucapan

Kata yang salah eja (*Misspelled*) atau disalahgunakan dapat menimbulkan masalah untuk analisis teks. Aplikasi koreksi otomatis dan koreksi tata bahasa dapat menangani kesalahan umum, tetapi tidak selalu memahami maksud penulis.

Dengan bahasa lisan, salah pengucapan, aksen berbeda, gagap, dll., bisa sulit dipahami oleh mesin. Namun, seiring berkembangnya basis data bahasa dan asisten cerdas dilatih oleh pengguna masing-masing, masalah ini dapat diminimalkan.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Bahasa sehari-hari dan bahasa *gaul* (slang)

Ungkapan informal, ekspresi, idiom, dan istilah khusus budaya menghadirkan sejumlah masalah untuk *NLP* – terutama untuk model yang ditujukan untuk penggunaan luas. Karena sebagai bahasa formal, bahasa sehari-hari mungkin tidak memiliki "definisi kamus" / *dictionary definition* sama sekali, dan ungkapan ini bahkan mungkin memiliki arti yang berbeda di wilayah geografis yang berbeda. Selain itu, bahasa *gaul* (slang) budaya terus berubah dan berkembang, sehingga kata-kata baru bermunculan setiap hari.

Di sinilah pelatihan dan pembaruan model *custom* secara teratur dapat membantu, meskipun seringkali membutuhkan data yang cukup banyak.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Bahasa khusus domain

Bisnis dan industri yang berbeda sering kali menggunakan bahasa yang sangat berbeda. Model pemrosesan NLP yang diperlukan untuk perawatan kesehatan (*healthcare*), misalnya, akan sangat berbeda dari yang digunakan untuk memproses dokumen hukum. Namun saat ini, ada sejumlah alat analisis yang dilatih untuk bidang tertentu, tetapi industri yang sangat khusus mungkin perlu membuat atau melatih model mereka sendiri.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Bahasa sumber daya rendah

Aplikasi *machine learning NLP* sebagian besar dibuat untuk bahasa yang paling umum dan banyak digunakan. Dan sungguh menakjubkan betapa akuratnya sistem terjemahan. Namun, banyak bahasa, terutama yang dituturkan oleh orang-orang yang kurang memiliki akses ke teknologi sering diabaikan dan kurang diproses. Misalnya, menurut beberapa perkiraan, (bergantung pada bahasa vs. dialek) ada lebih dari 3.000 bahasa di Afrika saja. Tidak banyak data tentang banyak bahasa ini.

Namun, teknik baru, seperti transformer multibahasa (menggunakan *Google's BERT* “*Bidirectional Encoder Representations from Transformers*”) dan penyematan kalimat *multibahasa* bertujuan untuk *mengidentifikasi* dan meningkatkan kesamaan universal yang ada di antara bahasa.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Kurangnya **penelitian** dan **pengembangan**

Machine Learning membutuhkan BANYAK data untuk berfungsi hingga batas luarnya – miliaran keping data pelatihan. Semakin banyak data model *NLP* dilatih, semakin pintar mereka jadinya. Yang mengatakan, data (dan bahasa manusia!) Hanya berkembang dari hari ke hari, begitu pula teknik *Machine Learning* baru dan algoritme khusus. Semua masalah di atas akan membutuhkan lebih banyak penelitian dan teknik baru untuk memperbaikinya.

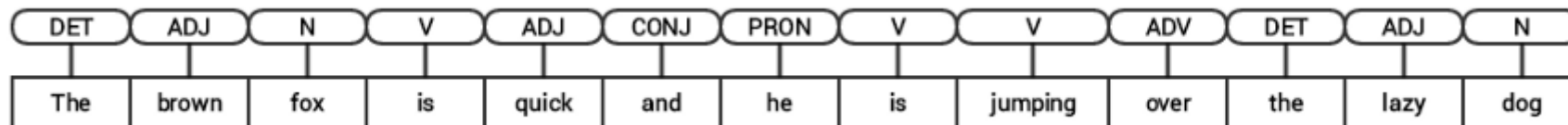
Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Aturan Sintaksis (*Syntactic Rules*)

Pengetahuan tentang **struktur** dan **sintaksis** bahasa sering membantu, dan beberapa teknik penguraian umum untuk memahami sintaksis teks dijelaskan di bawah ini.

- ✓ **POS Tagging** : *Parts of speech* (POS) adalah kategori leksikal spesifik tempat kata-kata diberikan, berdasarkan peran dan konteksnya dalam kalimat tertentu.



<https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7>

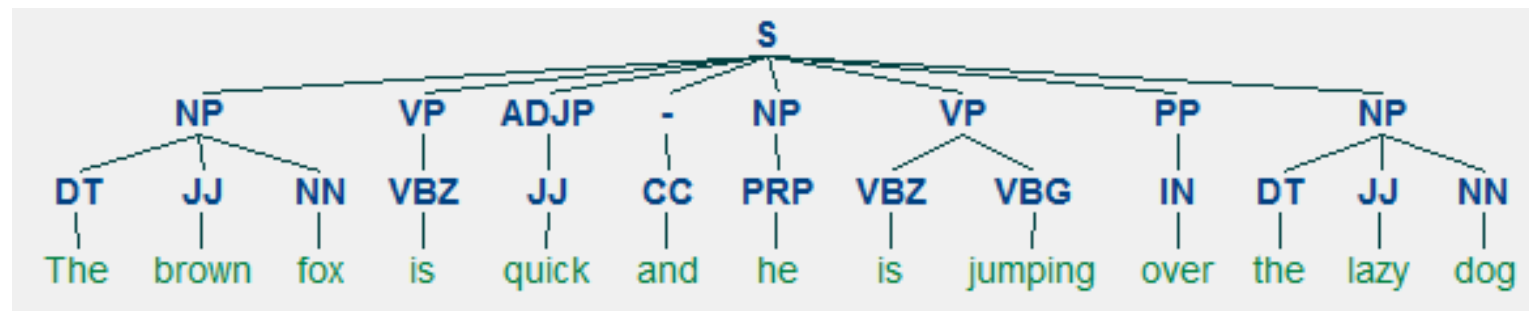
Misalnya, dalam kalimat di atas, " **he brown fox is quick, and he is jumping over the lazy dog**", singkatannya menunjukkan bagian-bagian pembicaraan berikut; **DET**: *Dependency tag*, **ADJ**: *Adjective*, **N**: *Noun*, **V**: *Verb*, **CONJ**: *Conjunction* (coordinating), **PRON**: *Pronoun*, **ADV**: *Adverb*.

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Aturan Sintaksis (*Syntactic Rules*)

- ✓ **Shallow Parsing/Chunking**: *Shallow parsing*, juga dikenal sebagai *chunking*, adalah metode menganalisis struktur kalimat dan memecahnya menjadi konstituen terkecil, yang biasanya berupa token seperti kata, dan kemudian mengelompokkannya menjadi frasa.



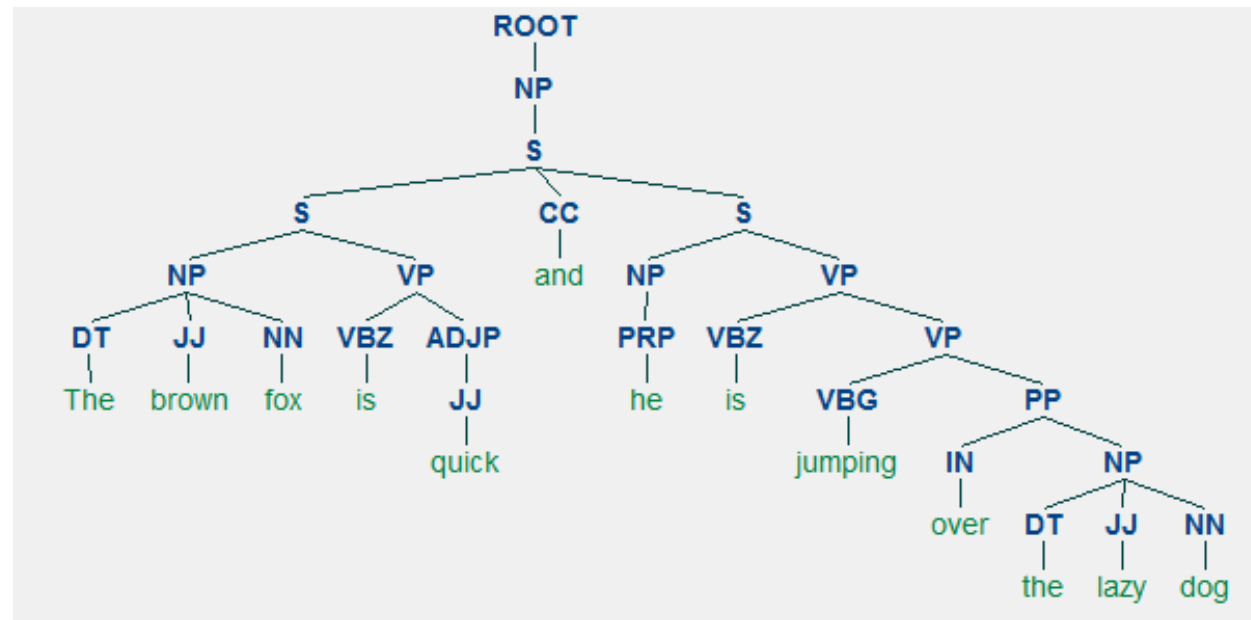
<https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7>

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Aturan Sintaksis (*Syntactic Rules*)

- ✓ **Constituency Parsing** : Parsing konstituensi bertujuan untuk mengekstrak pohon parse berbasis konstituensi dari sebuah kalimat. Pohon parse mewakili struktur sintaksisnya sesuai dengan tata bahasa struktur frase.



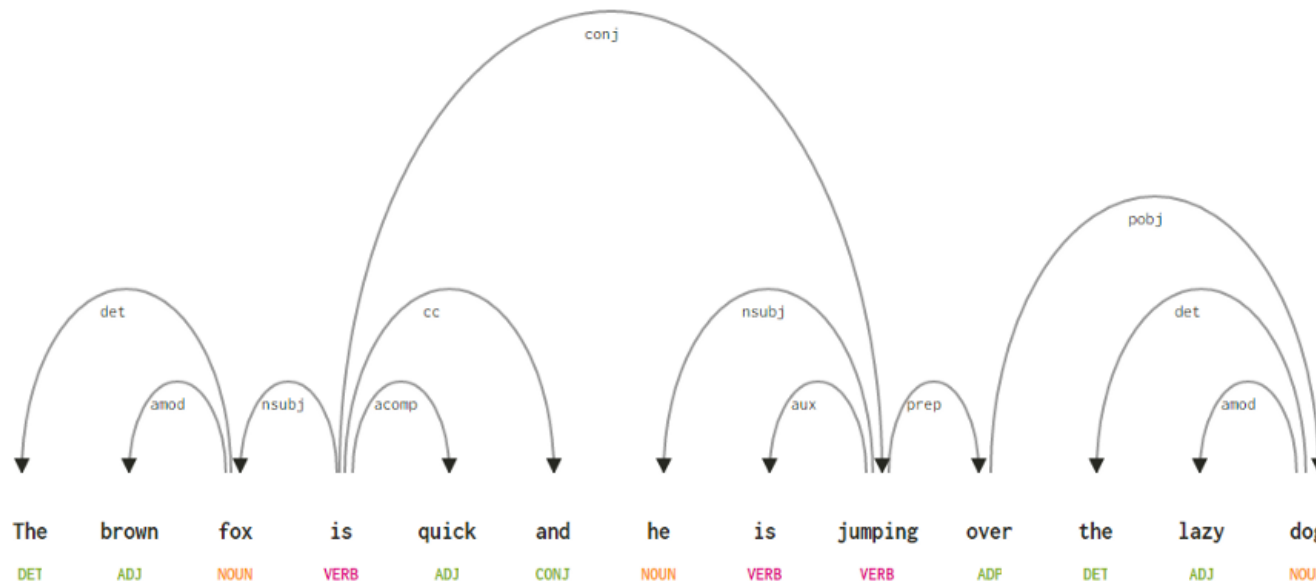
<https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7>

Tantangan Utama *Natural Language Processing*

<https://monkeylearn.com/blog/natural-language-processing-challenges/>

❑ Aturan Sintaksis (*Syntactic Rules*)

- ✓ *Dependency Parsing* : *Dependency parsing* adalah tugas menganalisis struktur tata bahasa kalimat dengan membangun hubungan antara kata-kata "*head*" dan kata-kata yang memodifikasi kepala tersebut.



<https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7>

Referensi :

- https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html
- <https://www.javatpoint.com/nlp>
- <https://thinkml.ai/what-are-the-natural-language-processing-challenges-and-how-to-fix-them/>
- <https://www.ibm.com/id-en/topics/natural-language-processing>
- <https://thinkml.ai/what-are-the-natural-language-processing-challenges-and-how-to-fix-them/>
- <https://www.hostcomm.co.uk/blog/2019/evolution-natural-language-infographic>
- <https://datasciencedojo.com/blog/natural-language-processing-applications/>
- <https://www.mvtvwireless.com/why-are-my-emails-going-into-spam/>
- <https://www.askpython.com/python/sentiment-analysis-using-python>
- <https://monkeylearn.com/blog/natural-language-processing-challenges/>
- <https://towardsai.net/p/nlp/natural-language-processing-concepts-and-workflow-48083d2e3ce7>
- <https://www.nltk.org/>