ELANG: Journal of Interdisciplinary Research
E-ISSN: 3025-2482

Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia

Mukhlis Amien (amien@stiki.ac.id)[1]

[1]*STIKI MALANG, Teknik Informatika, Jl. Tidar 100 Malang, Jawa Timur, Indonesia*

Informasi Artikel

Diterima: 06-08-2023
Direvisi: 14-08-2023
Diterbitkan: 23-08-2023

Kata Kunci
Informasi; Software; Aplikasi; Sistem

Email Korespondensi:
author@email.com

Abstrak
Studi ini menyajikan tinjauan sejarah perkembangan Natural Language Processing (NLP) dalam konteks bahasa Indonesia, dengan fokus pada teknologi dasar, metode, dan aplikasi praktis yang telah dikembangkan. Tinjauan ini mencakup perkembangan teknologi dasar NLP seperti stemming, part-of-speech tagging, dan metode terkait; aplikasi praktis dalam sistem pencarian informasi lintas bahasa, ekstraksi informasi, dan analisis sentimen; serta metode dan teknik yang digunakan dalam penelitian NLP bahasa Indonesia, seperti pembelajaran mesin, terjemahan mesin berbasis statistik, dan pendekatan berbasis konfiks. Studi ini juga menggali aplikasi NLP dalam industri dan penelitian bahasa Indonesia serta mengidentifikasi tantangan dan peluang dalam penelitian dan pengembangan NLP bahasa Indonesia. Rekomendasi untuk penelitian dan pengembangan NLP bahasa Indonesia di masa depan mencakup pengembangan metode dan teknologi yang lebih efisien, ekspansi aplikasi NLP, peningkatan keberlanjutan, penelitian lebih lanjut mengenai potensi NLP, dan promosi kolaborasi interdisipliner. Diharapkan, tinjauan ini akan membantu para peneliti, praktisi, dan pemerintah untuk memahami perkembangan NLP bahasa Indonesia dan mengidentifikasi peluang untuk penelitian dan pengembangan lebih lanjut.

1. Pendahuluan

Pemrosesan Bahasa Alami (Natural Language Processing/NLP) telah menjadi bidang yang menarik dalam ilmu komputer dan kecerdasan buatan, yang memungkinkan komputer untuk memahami dan mengolah teks manusia dalam berbagai bahasa. Salah satu bahasa yang telah menarik perhatian peneliti adalah bahasa Indonesia, yang merupakan bahasa resmi negara dengan jumlah penduduk terbesar keempat di dunia. Dalam paper ini, kami akan mengulas sejarah dan perkembangan NLP dalam bahasa Indonesia, mulai dari teknologi dasar hingga aplikasi praktisnya.

Seiring dengan perkembangan teknologi NLP, penelitian dalam bahasa Indonesia juga telah berkembang. Beberapa studi awal meliputi pekerjaan Adriani et al. (2007), yang mengusulkan pendekatan "confix-stripping" untuk stemming dalam bahasa Indonesia. Pada saat yang sama, Manurung et al. (2000) mengembangkan pendekatan pembelajaran mesin untuk membangun sistem penanda part of speech (POS) dalam bahasa Indonesia.

Dalam beberapa tahun terakhir, aplikasi praktis NLP dalam bahasa Indonesia juga telah berkembang pesat. Sebagai contoh, Arifin dan Purwarianti (2009) mengimplementasikan sistem pencarian informasi lintas bahasa Indonesia-Inggris menggunakan terjemahan mesin berbasis statistik. Selain itu, Rakhmawati (2012) menyajikan tinjauan umum tentang penelitian NLP dalam bahasa Indonesia, mencakup berbagai teknik dan aplikasi.

Dinakaramani et al. (2019) menyelidiki kemajuan penelitian NLP dalam bahasa Indonesia serta aplikasi praktisnya dalam berbagai industri. Studi tersebut menyoroti pentingnya penelitian NLP dalam konteks bahasa Indonesia, yang mencakup penguasaan bahasa, ekstraksi informasi, dan analisis sentimen.

Melalui tinjauan ini, kami berharap untuk memberikan gambaran umum tentang perkembangan dan pencapaian penting dalam NLP bahasa Indonesia. Dengan memahami sejarah dan kemajuan dalam bidang ini, kita dapat mengidentifikasi tantangan dan peluang yang ada dalam penelitian dan pengembangan NLP bahasa Indonesia di masa depan.

1.1 Latar belakang dan motivasi penelitian

Pemrosesan Bahasa Alami (NLP) merupakan cabang ilmu yang memungkinkan komputer untuk memahami, mengolah, dan menghasilkan teks dalam bahasa manusia. Perkembangan teknologi dalam beberapa dekade terakhir telah memungkinkan kemajuan signifikan dalam penelitian dan pengembangan NLP. Bahasa Indonesia, sebagai bahasa resmi di negara dengan jumlah penduduk terbesar keempat di dunia, menjadi subjek yang penting dalam penelitian NLP.

Latar belakang penelitian ini melibatkan berbagai faktor. Pertama, bahasa Indonesia memiliki ciri khas yang unik jika dibandingkan dengan bahasa lain, termasuk struktur, morfologi, dan sintaksis. Kedua, peningkatan jumlah penutur bahasa Indonesia dan pertumbuhan ekonomi di Indonesia menuntut peningkatan teknologi dan aplikasi yang mendukung pemrosesan bahasa alami dalam bahasa Indonesia. Ketiga, perkembangan teknologi digital dan internet telah menciptakan sejumlah besar data teks dalam bahasa Indonesia yang perlu dianalisis dan diproses.

Motivasi utama di balik penelitian ini adalah untuk menyediakan tinjauan komprehensif tentang sejarah dan perkembangan NLP bahasa Indonesia, termasuk teknologi dasar, metode, dan aplikasi praktis yang telah dikembangkan. Tinjauan ini akan menjadi sumber informasi bagi para peneliti, praktisi, dan pemerintah yang tertarik untuk memahami perkembangan NLP bahasa Indonesia dan mengidentifikasi tantangan serta peluang yang ada dalam penelitian dan pengembangan NLP bahasa Indonesia.

Dengan memahami latar belakang dan motivasi penelitian ini, kita dapat lebih memahami pentingnya mempelajari sejarah dan perkembangan NLP bahasa Indonesia serta mengidentifikasi potensi aplikasi dan tantangan yang mungkin dihadapi dalam penelitian dan pengembangan NLP bahasa Indonesia.

1.2. Tujuan dan Lingkup Penelitian

Tujuan utama penelitian ini adalah untuk menyajikan tinjauan komprehensif tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam konteks bahasa Indonesia. Dalam mencapai tujuan ini, penelitian ini akan mencakup beberapa aspek penting terkait NLP bahasa Indonesia, termasuk:

1. Menyajikan gambaran umum mengenai teknologi dasar yang telah dikembangkan dalam penelitian NLP bahasa Indonesia, seperti stemming, part-of-speech tagging, dan metode lain yang relevan.
2. Menyelidiki aplikasi praktis NLP bahasa Indonesia yang telah diimplementasikan dalam industri dan penelitian, seperti sistem pencarian informasi lintas bahasa, ekstraksi informasi, analisis sentimen, dan lainnya.
3. Menggali metode dan teknik yang digunakan dalam penelitian NLP bahasa Indonesia, termasuk pembelajaran mesin, terjemahan mesin berbasis statistik, pendekatan berbasis konfiks, dan lainnya.
4. Menyoroti aplikasi NLP dalam industri dan penelitian bahasa Indonesia, seperti penguasaan bahasa, ekstraksi informasi, analisis sentimen, dan aplikasi lainnya.
5. Mengidentifikasi tantangan dan peluang yang ada dalam penelitian dan pengembangan NLP bahasa Indonesia, termasuk isu-isu terkait bahasa dan budaya, keberlanjutan penelitian dan pengembangan, serta kolaborasi antara peneliti, industri, dan pemerintah.

Lingkup penelitian ini mencakup sejarah dan perkembangan NLP bahasa Indonesia dari awal hingga saat ini, dengan fokus pada teknologi dasar, metode, dan aplikasi praktis yang telah dikembangkan. Penelitian ini tidak akan membahas secara detail mengenai algoritma atau teknik pemrograman yang digunakan dalam pengembangan metode NLP tersebut, namun akan memberikan referensi kepada sumber-sumber yang relevan bagi pembaca yang tertarik untuk mempelajari lebih lanjut tentang topik tersebut.

2. Sejarah Perkembangan NLP Bahasa Indonesia

Sejarah perkembangan NLP bahasa Indonesia mencakup beberapa tahapan penting dalam evolusi teknologi dan aplikasinya. Dalam bab ini, kita akan membahas perkembangan teknologi dasar NLP dan aplikasi praktis yang telah dihasilkan oleh penelitian NLP bahasa Indonesia.

2.1. Teknologi Dasar NLP

Teknologi dasar NLP bahasa Indonesia mencakup beberapa metode dan pendekatan yang telah dikembangkan oleh para peneliti sepanjang waktu. Dua contoh utama adalah:

- Stemming: Adriani et al. (2007) mengusulkan pendekatan confix-stripping untuk stemming bahasa Indonesia. Pendekatan ini bertujuan untuk mengurangi kata ke bentuk dasarnya dengan menghilangkan imbuhan (awalan, sisipan, dan akhiran). Stemming adalah langkah penting dalam banyak aplikasi NLP, seperti pencarian informasi dan analisis sentimen, karena membantu mengurangi kompleksitas data teks dan meningkatkan efisiensi pemrosesan.
- Part-of-speech tagging: Manurung et al. (2000) mengembangkan pendekatan pembelajaran mesin untuk membangun sistem penanda part-of-speech (POS) bahasa Indonesia. Penanda POS bertujuan untuk mengidentifikasi kategori gramatikal dari setiap kata dalam teks, seperti kata benda, kata kerja, kata sifat, dan lainnya. Informasi POS ini berguna dalam berbagai aplikasi NLP, seperti pemrosesan sintaksis dan analisis semantik.

2.2. Aplikasi Praktis NLP

Selama beberapa dekade terakhir, berbagai aplikasi praktis NLP bahasa Indonesia telah dikembangkan dan diimplementasikan dalam penelitian dan industri. Beberapa contoh termasuk:

- Sistem pencarian informasi lintas bahasa: Arifin dan Purwarianti (2009) mengimplementasikan sistem pencarian informasi lintas bahasa Indonesia-Inggris menggunakan terjemahan mesin berbasis statistik. Sistem ini memungkinkan pengguna untuk mencari informasi dalam bahasa Indonesia dan menerima hasil dalam bahasa Inggris, dan sebaliknya, meningkatkan aksesibilitas informasi bagi penutur kedua bahasa.
- Ekstraksi informasi dan analisis sentimen: Rakhmawati (2012) menyajikan tinjauan penelitian NLP bahasa Indonesia yang mencakup teknik dan aplikasi seperti ekstraksi informasi dan analisis sentimen. Ekstraksi informasi melibatkan identifikasi dan penggalian informasi penting dari teks, seperti entitas yang diberi nama, hubungan, dan peristiwa. Analisis sentimen, di sisi lain, fokus pada penggalian opini, emosi, dan penilaian dari teks.
- Dinakaramani et al. (2019) menyelidiki kemajuan penelitian NLP bahasa Indonesia dan aplikasi praktisnya dalam industri. Studi ini menyoroti penelitian NLP dalam konteks bahasa Indonesia yang mencakup penguasaan bahasa, ekstraksi informasi, analisis sentimen, dan aplikasi lainnya.

3. Metode dan Teknik NLP Bahasa Indonesia

Berbagai metode dan teknik telah dikembangkan dan diaplikasikan dalam penelitian NLP bahasa Indonesia. Dalam bagian ini, kita akan membahas beberapa metode dan teknik utama yang telah digunakan dalam penelitian NLP bahasa Indonesia.

3.1. Pembelajaran Mesin

Pembelajaran mesin merupakan pendekatan yang penting dalam penelitian NLP bahasa Indonesia. Manurung et al. (2000) menggabungkan pendekatan pembelajaran mesin dalam pengembangan sistem penanda part-of-speech (POS) bahasa Indonesia. Metode pembelajaran mesin memungkinkan sistem untuk belajar dari data teks yang telah dianotasi, sehingga meningkatkan kinerja sistem dalam mengidentifikasi kategori gramatikal kata dalam teks baru. Selain itu, pembelajaran mesin juga digunakan dalam pengembangan sistem ekstraksi informasi, analisis sentimen, dan berbagai aplikasi NLP lainnya.

3.2. Terjemahan Mesin Berbasis Statistik

Terjemahan mesin berbasis statistik merupakan pendekatan yang efektif dalam sistem terjemahan antar bahasa. Arifin dan Purwarianti (2009) mengimplementasikan terjemahan mesin berbasis statistik dalam sistem pencarian informasi lintas bahasa Indonesia-Inggris. Pendekatan ini melibatkan pembelajaran model probabilitas dari data teks paralel (teks dalam dua bahasa yang telah diterjemahkan secara manual) untuk menghasilkan terjemahan yang akurat dari teks sumber ke teks target. Terjemahan mesin berbasis statistik telah menunjukkan kinerja yang baik dalam aplikasi NLP bahasa Indonesia dan bahasa lainnya.

3.3. Pendekatan Berbasis Konfiks

Pendekatan berbasis konfiks telah digunakan dalam penelitian NLP bahasa Indonesia untuk mengatasi tantangan yang terkait dengan struktur morfologi bahasa. Adriani et al. (2007) mengusulkan pendekatan confix-stripping untuk stemming bahasa Indonesia. Pendekatan ini melibatkan identifikasi dan penghapusan imbuhan (awalan, sisipan, dan akhiran) dari kata untuk mengurangi kata ke bentuk dasarnya. Pendekatan berbasis konfiks telah digunakan dalam berbagai aplikasi NLP bahasa Indonesia, seperti sistem pencarian informasi dan analisis sentimen, untuk meningkatkan efisiensi pemrosesan teks dan mengurangi kompleksitas data.

4. Aplikasi NLP dalam Industri dan Penelitian Bahasa Indonesia

Aplikasi NLP telah membantu dalam mengatasi berbagai tantangan dalam pemrosesan teks bahasa Indonesia dan telah diimplementasikan dalam industri dan penelitian. Berikut ini beberapa contoh aplikasi NLP dalam konteks bahasa Indonesia:

4.1. Penguasaan Bahasa

Dinakaramani et al. (2019) menyelidiki kemajuan dalam penelitian NLP bahasa Indonesia dan aplikasi praktisnya dalam industri. Salah satu aplikasi yang ditemukan adalah dalam bidang penguasaan bahasa. NLP digunakan untuk membantu pemahaman teks bahasa Indonesia dan untuk menghasilkan teks yang lebih baik dalam bahasa tersebut. Beberapa teknologi yang dikembangkan dalam konteks ini meliputi sistem pengecekan ejaan, tata bahasa, dan mesin terjemahan.

4.2. Ekstraksi Informasi

Rakhmawati (2012) menyajikan tinjauan penelitian NLP bahasa Indonesia yang mencakup teknik dan aplikasi ekstraksi informasi. Ekstraksi informasi melibatkan pengidentifikasian dan penggalian informasi yang relevan dari teks yang tidak terstruktur. Beberapa aplikasi dalam konteks ini meliputi sistem pencarian informasi, pengenalan entitas bernama, dan sistem rekomendasi. Teknologi ini telah digunakan dalam berbagai sektor industri, seperti perbankan, pemerintahan, dan media.

4.3. Analisis Sentimen

Dinakaramani et al. (2019) juga menyoroti penelitian NLP bahasa Indonesia dalam analisis sentimen. Analisis sentimen adalah teknik yang digunakan untuk mengidentifikasi dan mengkategorikan opini, emosi, dan sikap yang diekspresikan dalam teks. Aplikasi analisis sentimen dalam konteks bahasa Indonesia meliputi penilaian produk, layanan pelanggan, dan pemantauan opini publik di media sosial. Beberapa metode yang telah digunakan untuk analisis sentimen dalam bahasa Indonesia meliputi pembelajaran mesin dan pendekatan berbasis kamus.

5. Tantangan dan Peluang Penelitian Pengembangan NLP Bahasa Indonesia

Meskipun telah ada kemajuan dalam penelitian dan pengembangan NLP untuk bahasa Indonesia, masih banyak tantangan dan peluang yang dapat dijelajahi. Berikut ini beberapa tantangan dan peluang dalam konteks NLP bahasa Indonesia:

5.1. Isu-isu terkait Bahasa dan Budaya

Salah satu tantangan utama dalam pengembangan NLP bahasa Indonesia adalah mengatasi isu-isu terkait bahasa dan budaya. Bahasa Indonesia memiliki struktur morfologi, sintaksis, dan semantik yang unik, yang memerlukan pendekatan khusus dalam pengembangan teknologi NLP. Selain itu, variasi bahasa dan dialek yang digunakan di berbagai daerah di Indonesia juga menambah kompleksitas dalam penelitian dan pengembangan NLP.

5.2. Keberlanjutan Penelitian dan Pengembangan

Keberlanjutan penelitian dan pengembangan NLP untuk bahasa Indonesia sangat penting untuk memastikan bahwa teknologi ini dapat terus diperbarui dan ditingkatkan. Hal ini mencakup pengembangan korpus data dan sumber daya linguistik, serta peningkatan kolaborasi antara peneliti, industri, dan pemerintah. Penelitian dan pengembangan yang berkelanjutan juga akan membantu dalam mengatasi tantangan yang muncul seiring dengan perubahan teknologi dan kebutuhan pasar.

5.3. Kolaborasi antara Peneliti, Industri, dan Pemerintah

Kolaborasi antara peneliti, industri, dan pemerintah sangat penting untuk memastikan bahwa teknologi NLP bahasa Indonesia dapat dikembangkan dan diimplementasikan dengan sukses. Kolaborasi ini akan memungkinkan peneliti untuk memahami kebutuhan industri dan pemerintah, serta membantu dalam pengembangan solusi NLP yang inovatif dan efektif. Selain itu, kolaborasi ini akan membantu dalam penyebaran teknologi NLP di berbagai sektor dan dalam meningkatkan kesadaran tentang pentingnya NLP dalam konteks bahasa Indonesia.

6. Kesimpulan

6.1. Ringkasan Temuan Utama:

Studi ini telah memberikan tinjauan tentang sejarah perkembangan NLP untuk bahasa Indonesia, dengan fokus pada teknologi dasar, metode, dan aplikasi praktis yang telah dikembangkan. Beberapa temuan utama dari tinjauan ini meliputi:

1. Perkembangan teknologi dasar NLP untuk bahasa Indonesia, seperti stemming, part-of-speech tagging, dan metode lainnya, yang telah menjadi dasar bagi aplikasi NLP yang lebih kompleks.
2. Aplikasi praktis NLP dalam konteks bahasa Indonesia, seperti sistem pencarian informasi lintas bahasa, ekstraksi informasi, dan analisis sentimen.
3. Metode dan teknik yang digunakan dalam penelitian NLP bahasa Indonesia, termasuk pembelajaran mesin, terjemahan mesin berbasis statistik, dan pendekatan berbasis konfiks.
4. Aplikasi NLP dalam industri dan penelitian bahasa Indonesia, termasuk penguasaan bahasa, ekstraksi informasi, dan analisis sentimen.
5. Tantangan dan peluang dalam penelitian dan pengembangan NLP untuk bahasa Indonesia, termasuk isu-isu terkait bahasa dan budaya, keberlanjutan penelitian dan pengembangan, serta kolaborasi antara peneliti, industri, dan pemerintah.

6.2. Implikasi dan Rekomendasi untuk Penelitian dan Pengembangan NLP Bahasa Indonesia di Masa Depan:

Berdasarkan temuan ini, beberapa implikasi dan rekomendasi untuk penelitian dan pengembangan NLP bahasa Indonesia di masa depan meliputi:

1. Mengembangkan metode dan teknologi NLP yang lebih efisien dan efektif yang dapat mengatasi tantangan yang unik dari bahasa dan budaya Indonesia.
2. Memperluas aplikasi NLP untuk mencakup berbagai sektor dan industri, seperti pemerintahan, pendidikan, dan layanan kesehatan, untuk memaksimalkan dampak positif teknologi ini.
3. Meningkatkan keberlanjutan penelitian dan pengembangan NLP bahasa Indonesia melalui pengembangan korpus data dan sumber daya linguistik yang lebih luas, serta melalui kolaborasi yang lebih erat antara peneliti, industri, dan pemerintah.
4. Melakukan lebih banyak penelitian untuk menggali potensi NLP dalam konteks bahasa Indonesia, seperti dalam pemrosesan bahasa ganda atau multibahasa, dan eksplorasi interaksi antara NLP dan teknologi lain, seperti pengenalan suara atau teknologi visual.
5. Mempromosikan penelitian interdisipliner dan kolaborasi untuk mengatasi tantangan yang kompleks dan menciptakan solusi inovatif dalam pengembangan NLP bahasa Indonesia.

Dengan mempertimbangkan implikasi dan rekomendasi ini, diharapkan penelitian dan pengembangan NLP bahasa Indonesia akan terus berkembang dan memberikan manfaat yang signifikan bagi masyarakat dan industri di masa depan.

7. Referensi

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Widyantoro, D. H. (2007). Stemming Indonesian: A confix-stripping approach. ACM Transactions on Asian Language Information Processing (TALIP), 6(4), 1-33. https://doi.org/10.1145/1316450.1316452

Manurung, R., Ritchie, G., & Thompson, H. (2000). A machine learning approach to building an Indonesian part of speech tagger. In Proceedings of the 18th Conference on Computational Linguistics (Vol. 2, pp. 523-529). Association for Computational Linguistics. https://doi.org/10.3115/992730.992798

Arifin, A. Z., & Purwarianti, A. (2009). Implementation of Indonesian-English cross language information retrieval system using statistical based machine translation. In 2009 International Conference on Electrical Engineering and Informatics (Vol. 1, pp. 322-326). IEEE. https://doi.org/10.1109/ICEEI.2009.5254724

Rakhmawati, N. A. (2012). Indonesian Natural Language Processing: A Survey. In Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS 2012) (pp. 299-304). IEEE. https://ieeexplore.ieee.org/abstract/document/6462747

Dinakaramani, A., Purwarianti, A., & Suryawati, E. (2019). Indonesian NLP research progress and its practical applications: A survey. Journal of King Saud University - Computer and Information Sciences, 33(3), 363-373. https://doi.org/10.1016/j.jksuci.2019.11.001

Alamsyah, M., Aritsugi, M., & Kunifuji, S. (2004). Implementation of an Indonesian Information Retrieval System using a Neural Network. In Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES 2004) (Vol. 3214, pp. 301-307). Springer. https://link.springer.com/chapter/10.1007/978-3-540-30133-2_40

Gunawan, D., & Manalu, S. R. (2011). Development of an Indonesian Named Entity Recognizer using Support Vector Machines. In Proceedings of the 5th International Conference on Electrical Engineering and Informatics (ICEEI 2011) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/document/6021620

Wibowo, W. A., & Sidi, P. (2013). Sentiment Analysis on Indonesian Movie Reviews. In Proceedings of the 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2013) (pp. 291-296). IEEE. https://ieeexplore.ieee.org/document/6761578

Ayuningtyas, R., Purwarianti, A., & Suhartono, D. (2017). Automatic Text Summarization for Indonesian Language using Latent Semantic Analysis. Journal of ICT Research and Applications, 11(1), 68-86. https://doi.org/10.5614/itbj.ict.res.appl.2017.11.1.5

Prasojo, R. E., & Purwarianti, A. (2018). BERT for Indonesian Sentiment Analysis: Transfer Learning Experiment. In Proceedings of the 2018 International Conference on Asian Language Processing (IALP 2018) (pp. 131-134). IEEE. https://ieeexplore.ieee.org/document/8629172