Tinjauan Literatur : Identifikasi Dialek Dengan *Deep Learning*

Rheza Daffa Pamungkas
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl.Kaliurang KM 14.5, Sleman, Yogyakarta, Indonesia
17523182@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl.Kaliurang KM 14.5, Sleman, Yogyakarta, Indonesia
Fathan@uii.ac.id

Abstract*— Identifikasi dialek merupakan sebuah sub-bagian dari identifikasi bahasa yang lebih fokus dalam bagian dialek yang memiliki tingkat kemiripan dalam satu bahasa. Identifikasi dialek diperuntukkan untuk membedakan dan mengklasifikasikan dialek kedalam kelas yang diinginkan. Melalui penelitian ini, peneliti mengkaji literatur pada tahun 2016 hingga 2020 mengenai identifikasi dialek yang didapatkan dari Google Scholar dengan kata kunci "Dialect Identification". Melalui tinjauan literatur ini didapati pada umumnya identifikasi dialek dilakukan dengan dua pendekatan yaitu pendekatan machine learning dengan model SVM sebagai model machine learning yang paling banyak digunakan dan pendekatan deep learning dengan model BiLSTM sebagai model deep learning yang banyak digunakan. Model SVM digunakan karena SVM membantu dalam melakukan pengklasifikasian dialek dengan menunjukkan performa yang baik dalam beberapa kasus mengenai pengklasifikasian idenfikasi bahasa termasuk identifikasi dialek dan model BiLSTM digunakan karena dapat memahami dan mengambil perspektif dari kata sebelumnya dan kata yang setelahnya, sehingga proses pembelajaran akan semakin bertmbah yang berdampak pada model sehingga lebih memahami konteks pada dialek tersebut. Model BiLSTM akan bermanfaat dalam pengklasifikasian dialek apabila memiliki akses terhadap kedua informasi tersebut.Tinjauan literatur ini memiliki tujuan untuk digunakan oleh penelitian lainnya sebagai referensi pengembangan penelitian identifikasi dialek.

Keywords—identifikasi dialek, machine learning, deep learning, tinjauan literatur

I. PENDAHULUAN

Sebuah dialek yang diucapkan oleh bahasa mana saja merupakan representasi dari sebuah pola pengucapan yang biasanya diucapkan oleh penutur asli dengan karakteristik unik di daerah tertentu. Dialek sendiri mayoritas besar penuturnya terpengaruhi oleh daerah asli maupun daerah yang ditempatinya dengan berbagai penyebab diantaranya, lokasi geografis, tingkat sosial masyarakat dan juga budaya[1]. Maka identifikasi dialek ternilai lebih susah dibandingkan dengan identifikasi bahasa walaupun identifikasi dialek termasuk ke dalam sub-bagian dari identifikasi bahasa[2]. Hal ini dikarenakan tugas dari identifikasi dialek adalah untuk dapat membedakan sebuah dialek yang biasanya memiliki tingkat kemiripan dengan dialek lainnya dalam bahasa yang sama.

Negara Indonesia memiliki kurang lebih 742 bahasa daerah dengan 737 diantaranya masih hidup atau masih digunakan oleh penutur asli daerah tersebut.Dari sekian banyak bahasa daerah yang ada di Indonesia, Bahasa Jawa memiliki jumlah penutur yang paling banyak di Indonesia dengan diperkirakan mampu mencapai 75 juta penutur[3] dikarenakan penuturnya tidak hanya berasal dari Indonesia tetapi juga dari luar negeri. Maka dari itu Bahasa Jawa dapat menduduki peringkat 11 di dunia jika dilihat dari penuturnya dan digunakan oleh 5 negara lainnya selain Indonesia[4]. Bahasa jawa sendiri dapat dikategorikan kedalam 3 kategori yaitu bahasa jawa standar yang biasanya digunakan sekitar Jogja dan Solo, bahasa jawa timuran yang biasanya digunakan di daerah hampir seluruh daerah Jawa Timur kecuali Banyuwangi dan bahasa jawa Banyumas yang biasanya digunakan oleh daerah barat Kedu dan sekitar Banyumas[5].

Dengan begitu penelitian ini dilakukan dengan tujuan untuk membantu peneliti dalam menentukan metode *deep learning* yang tepat atau mendapatkan metode disarankan untuk melakukan identifikasi dialek bahasa jawa.Peneliti berharap melalui penelitian dapat membantu bidang NLP terkhususnya dalam bidang identifikasi dialek dalam menentukan model deep learning yang tepat untuk identifikasi dialek bahasa jawa pada teks dan dapat menemukan tren dalam bidang identifikasi dialek secara umum.

Berdasarkan penjelasan di atas maka, langkah pertama dalam melakukan penelirian tinjauan literatur dapat berupa sebuah pertanyaan penelitian. Pertanyaan-pertanyaaan ini dikemukakan oleh peneliti guna sebagai langkah awal atau sebagai arah untuk mendapatkan hasil yang diinginkan dalam penelitian ini. Pertanyaan yang dikemukakan oelh peneliti bersifat jelas dan singkat. Pertanyaan-pertanyaan tersebut dapat berupa bagaimana perkembangan tren untuk penelitian identifikasi dialek?, metode apa saja yang dapat digunakan dalam identifikasi dialek?, penelitian identifikasi dialek lebih banyak menggunakan dialek apa saja?, Bagaimana saran yang dapat diberikan untuk mambantu penelitian identifikasi dialek selanjutnya?. Pertanyaan tersebut akan dijawab oleh peneliti melalui penelitian tinjauan literatur ini.

II. METODELOGI PENELITIAN

A. *Pengumpulan Data*

Data yang dikumpulkan berupa literatur yang diperoleh melalui *Google Scholar*. Proses pengumpulan data dapat dibagi menjadi 3 tahapan yaitu, tahap pertama pengumpulan data literatur yaitu dengan memenuhi kriteria penelitian identifikasi dialek dengan tahun terbit mulai dari 2016 hingga 2020. Kata kunci yang digunakan adalah "*Dialect identification"*. Tahap kedua pengumpulan data literatur yaitu dengan memenuhi kriteria penelitian identifikasi dialek dengan tahun terbit mulai dari 2016 hingga 2020. Kata kunci yang digunakan adalah "*Chinese Dialect identification".* Tahap ketiga pengumpulan data litaratur dengan memenuhi kriteria penelitian identifikasi dialek dengan tahun terbit mulai

dari 2016 hingga 2020. Kata kunci yang digunakan adalah "*Dialect identification with deep learning".* Hasil total literatur yang diperoleh sebanyak 12 literatur yang dapat dilihat melalui *Table* 1 berikut.

TABLE 1 LITERATUR YANG DIPEROLEH DARI *GOOGLE SCHOLAR*

Refrensi	Tahun
[6]	2016
[7]	2016
[8]	2017
[9]	2018
[10]	2018
[11]	2018
[12]	2019
[13]	2019
[14]	2019
[15]	2019
[16]	2019
[17]	2020

III. PEMBAHASAN

A. *Identifikasi Dialek*

Dalam melakukan penelitian identifikasi dialek memerlukan beberapa tahapan yang harus dilalui untuk mendapatkan hasil keluaran yang diharapkan. Tahapan-tahapan tersebut dapat dilihat sebagai berikut.

1) Pengumpulan data

Data yang digunakan dalam penelitian identifikasi dialek dapat berupa data baru atau data yang sudah terbali dan yang sudah digunakan Data yang dikumpulkan dapat beragam dan dapat mengambil lebih dari satu sumber. Hal ini disesuaikan dengan kebutuhan sesuai dengan tujuan penelitian yang ingin dicapai. Dapat dilihat pada *Table* 2 yang menunjukkan data apa saja data dan dialek yang digunakan pada penelitian-penelitian sebelumnya.

Dapat dilihat berdasarkan paparan *Table 2* data yang digunakan bervariasi dan penggunaan dialek lebih banyak menggunakan Bahasa Arab atau *Arabic Dialect.* Dari 12 literatur yang peneliti ambil 9 diantaranya menggunakan dialek Bahasa Arab sebagai data yang digunakan dalam penelitian mereka dan 3 literatur sisanya menggunakan dialek Bahasa China atau mandarin dan Jerman.

2) *Preprocessing*

Tahapan *preprocessing* digunakan untuk mengolah data mentah untuk dirubah menjadi data yang siap diolah atau data yang siap untuk digunakan pada proses selanjutnya. Tahapan ini sendiri bermacam-macam sesuai dengan kebutuhan yang diperlukan pada penelitian yang dilakukan. Salah satu contoh pada penelitian Elaraby, et al[9], pada penelitian tersebut preprocessing yang dilakukan salah satunya adalah *tokenization and normalization* diperuntukkan untuk data dengan spasi dan menghilangkan huruf *unicode* selain itu digunakna juga untuk menghilangkan kata-kata yang tidak menggunakan Bahasa Arab.

3) Ekstraksi Fitur

Tahapan ekstraksi fitur sangatlah memabntu dalam proses pengolahan data, dikarenakan tidak semua data dapat secara langsung diolah dengan menggunakan *deep learning* maupun *machine learning* terutama jika data tersebut berupa teks maka data tersebut harus diubah menjadi data numerik terlebih dahulu. Ekstraksi fitur juga berguna dalam mengambil informasi penting dari sebuah data untuk diolah pada proses selanjutnya. Ekstraksi fitur yang digunakan pada penelitian-penelitian sebelumnya dapat dilihat pada *Table* 3 berikut ini.

Berdasarkan dari *Table* 3 penggunaan ektrasi fitur pada penelitian identifikasi dialek banyak menggunakan ekstraksi fitur *n-gram* dan TF-IDF untuk metode *machine learning* seperti penelitian oleh [6][8][9][11-17] yang memiliki tingkatan berbeda-beda. Tingkatan yang digunakan oleh penelitian sebelumnya diantaranya tingkatan huruf, tingkatan kata, dan tingkatan kalimat. *N-gram* diguanakan sebagai model untuk membagi kumpulan *string* yang besar kedalam bagian yang lebih kecil dan TF-IDF digunkan untuk memberikan sebuah bobot kedalam suatu kata. Sedangkan ekstraksi fitur yang digunakan dalam metode *deep learning* ialah *embedding layer* seperti dalam penelitian [7][10][14][17] yang terbagi kedalam beberapa tingkatan seperti dalam ekstraksi fitur *n-gram* dan TF-IDF.

TABLE 2 LITERATUR BERDASARKAN *DATASET*

Referensi	Dataset	Dialect
[6]	Data menggunakan penelitian dari jurnal lain	*Arabic Dialect*
[7]	*Sub-task 2 of Arabic dialect*	*Arabic Dialect*
[8]	Berita dan Wikipedia	*Grester China Region* (GCR) atau mandarin
[9]	*Arabic online Commentary (AOC)*	*Arabic Dialect*
[10]	Data menggunakan penelitian dari jurnal sebelumnya	*Arabic Dialect*
[11]	*GDI organizer*	*German Dialect*
[12]	Berita	*China dan Taiwan manadarin dengan menggunakan simplified Chinese dan traditional chinese*
[13]	*MADAR*	*Arabic Dialect*
[14]	*MADAR (Multi Arabic Application and Resource)*	*Arabic Dialect*
[15]	*MADAR Travel Domain dataset*	*Arabic Dialect*
[16]	*Twitter*	*Arabic Dialect*
[17]	*Twitter*	*Arabic Dialect*

TABLE 3 METODE DANEKSTRAKSI FITUR YANG DIGUNAKAN

Referensi	Ekstraksi Fitur
[6]	*n-gram*
[7]	*Embedding Layer*
[8]	*Character level n-gram*
[9]	TF-IDF *n-gram*
[10]	*Word embedding*
[11]	TF-IDF *Character level n-gram* *Word level n-gram*
[12]	*n-gram*
[13]	TF-IDF
[14]	*Character* TF-IDF *Word* TF-IDF *Word embedding*
[15]	*n-gram*
[16]	*Character n-gram* *Word n-gram*
[17]	*TF-IDF* *Word n-gram* *Character n-gram* *Embedding Layer*

4) Pemodelan

Pada tahapan selanjutnya adalah tahap pembangunan model yang digunakan untuk melakukan identifikasi dialek. Pembangunan model yang dilakukan pada penelitian sebelumnya dibangun dari beberapa model *machine learning* atau *deep learning* sesuai dengan kebutuhan guna mencapai hasil keluaran yang diinginkan seperti dalam *Table 4*. Model SVM adalah model *machine learning* yang paling banyak digunakan oleh penelitian sebelumnya seperti dalam penelitian [6][8-11][15][16] dan dapat menghasilkan akurasi rata-rata diatas 50%. Menurut [8] model SVM digunakan karena model SVM membantu dalam melakukan pengklasifikasian dialek dengan menunjukkan performa yang baik dalam beberapa kasus mengenai pengklasifikasian idenfikasi bahasa termasuk identifikasi dialek. BiLSTM sebagai model *deep learning* paling banyak diguakan pada penelitian sebelumnya seperti dalam penelitian [9][10][12][13] dan dapat mengahasilkan hasil akurasi yang baik dengan rata-rata diatas 80%. Menurut [10] dalam identifikasi dialek model BiLSTM dapat memahami dan mengambil perspektif dari kata sebelumnya dan kata yang setelahnya, sehingga proses pembelajaran akan semakin bertmbah yang berdampak pada model sehingga lebih memahami konteks pada dialek tersebut. Sehingga model BiLSTM akan bermanfaat dalam pengklasifikasian dialek apabila memiliki akses terhadap kedua informasi tersebut.

Model yang telah dibuat diukur tingkat keberhasilan dan keakuratannya pada tahap evaluasi. Penilaian tahapan evaluasi dapat menggunakan akurasi atau F1-*score*. Akurasi digunakan disaat jumlah data yang digunakan memiliki niali yang setara sedangan disaat jumlah data tidak seimbang maka penggunaaan akurasi saja tidak dapat mempresentasikan tingkatan kesuksesan sebuah model maka perlu ditambahkan nilai F1-*Score* yang didapatkan dari 2 kali rata-rata dari *precision* dan *recall*. Hasil tertinggi menurut *Table 4* penelitian oleh [8] dengan metode *machine learning* SVM dapat mencapai nilai akurasi 97% dan penelitian [9] dengan metode *deep learning* BiLSTM dapat mencapai rata-rata akurasi 85.13%.

TABLE 4 METODE DAN HASIL DARI PENELITAN SEBELUMNYA

Referensi	Metode	Hasil
[6]	LR SVM *language models*	F1 Score: 49.46% *and* 51.32%
[7]	CNN	*Accurcay:* 0.8307
[8]	SVM	*Accuracy :* Untuk berita 82%,90% dan 97% Untuk Wikipedia : 60% dan 77%
[9]	SVM LR MNB CNN LSTM CLSTM BiLSTM BiGRU BiLSTM with attention	*Accuracy :* BiGRU : 87.23% pada *binary data* *Attention* BiLSTM : 87.81 % pada *three-way data* dan 82.45% pada *four-way data*
[10]	SVM Multi-input CNN CNN-BiLSTM	CNN with separate embedding (F1score : 0.5289), CNN-BiLSTM (F1scoren :0.4235) Binary CNN-BiLSTM (F1score :0.4339)
[11]	SVM	F1 *Score* : 0.6203
[12]	LR SVM MNB CNN RNN Bi-LSTM	F1 *Score* : 0.8530 0.8687
[13]	MNB *Random Forest Classifier* BiLSTM	F1 *Score*: 63.02%

	RNN BiGRU	
[14]	MNB *Baseline* LSTM	*F1 score* : 65.66%
[15]	SVM MNB BNB LR SGD PA PC	*F1 Score :* 62%
[16]	SVM	*Accuraxy* : 65%
[17]	MNB *Linear* SVC BNB SGD Ridge LR LSTM	Macro Average F-Scoe: 18.8% Accuracy: 36.54%

KESIMPULAN

Berdasarkan tinjauan literatur ini didapati 12 literatur yang didapatkan melalui *Google scholar* yang diambil dalam rentang waktu 2016 hingga 2020. Literatur yang telah dapat dikaji didapati bahwa perkembangan tren penelitian identifikasi dialek dalam kurun waktu 2016 hingga 2020 adalah penelitian dapat menggunakan 2 pendekatan yaitu pendekatan *machine learning* dan *deep learning* dan dapat dikombinasikan juga satu dengan yang lainnya. Penggunaan model SVM menjadi model *machine learning* yang paling banyak digunakan dan model BiLSTM menjadi model *deep learning* yang paling banyak digunakan.

Berdasarkan pembahasan diatas model SVM membantu dalam melakukan pengklasifikasian dialek dengan menunjukkan performa yang baik dalam beberapa kasus mengenai pengklasifikasian idenfikasi bahasa termasuk identifikasi dialek dan model BiLSTM dapat memahami dan mengambil perspektif dari kata sebelumnya dan kata yang setelahnya, sehingga proses pembelajaran akan semakin bertmbah yang berdampak pada model sehingga lebih memahami konteks pada dialek tersebut. Sehingga model BiLSTM akan bermanfaat dalam pengklasifikasian diaelk apabila memiliki akses terhadap kedua informasi tersebut.. Model tersebut juga dibantu dengan menggunakan beberapa ekstraksi fitur yang dapat membantu sebuah model untuk meningkatkan hasil keluaran seperti *n-gram*, TF-IDF, dan *word embedding*. Berdasarkan tinjauan literatur ini, identifikasi dialek yang menggunakan Bahasa Indonesia terkhususnya dalam dialek Bahasa Jawa masih sedikit dilakukan. Maka dari itu, perkembangan penelitian identifikasi dialek dengan menggunakan dialek Bahasa Jawa perlu dikembangkan lagi.

REFERENCES

[1] N. B. Chittaragi, A. Limaye, N. T. Chandana, B. Annappa, and S. G. Koolagudi, "Automatic text-independent Kannada dialect identification system," in *Advances in Intelligent Systems and Computing*, 2019, doi: 10.1007/978-981-13-3338-5_8.

[2] V. P. K, A. S, V. R, and S. KP, "A Deep Learning Approach for Similar Languages, Varieties and Dialects," -, 2019, [Online]. Available: http://arxiv.org/abs/1901.00297.

[3] L. S. Aji, S. Sugiharti, and M. Salimi, "ANALYSIS OF JAVANESE LANGUAGE VOCABULARY SKILL FOR ELEMENTARY SCHOOL STUDENTS IN KEBUMEN DISTRICT," *Soc. Humanit. Educ. Stud. Conf. Ser.*, 2019, doi: 10.20961/shes.v1i2.26876.

[4] A. P. Ardhana, D. E. Cahyani, and Winarno, "Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods," in *Journal of Physics: Conference Series*, 2019, doi: 10.1088/1742-6596/1306/1/012049.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] A. I. Fauzi and D. Puspitorini, "Dialect and Identity: A Case Study of Javanese Use in WhatsApp and Line," in *IOP Conference Series: Earth and Environmental Science*, 2018, doi: 10.1088/1755-1315/175/1/012111.

[6] S. Malmasi, M. Zampieri, N. Ljubeši, P. Nakov, A. Ali, and J. Tiedemann, "Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task," *Proc. Third Work. NLP Similar Lang. Var. Dialects*, 2016.

[7] Y. Belinkov and J. Glass, "A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects," 2016, [Online]. Available: http://arxiv.org/abs/1609.07568.

[8] F. Xu, M. Wang, and M. Li, "Sentence-level dialects identification in the Greater China region," *arXiv*. 2017, doi: 10.5121/ijnlc.2016.5602

[9] M. Elaraby and M. Abdul-Mageed, "Deep Models for Arabic Dialect Identification on Benchmarked Data," *Proc. Fifth Work. NLP Similar Lang. Var. Dialects St. Fe, New Mex. USA*, 2018.

[10] E. Michon, M. Q. Pham, J. Crego, and J. Senellart, "Neural Network Architectures for Arabic Dialect Identification," *Proc. ofthe Fifth Work. NLP Similar Lang. Var. Dialects*, 2018.

[11] A. M. Ciobanu, S. Malmasi, and L. P. Dinu, "German dialect identification using classifier ensembles," *arXiv*. 2018.

[12] L. Yang and Y. Xiang, "Naive {B}ayes and {B}i{LSTM} Ensemble for Discriminating between Mainland and Taiwan Variation of {M}andarin {C}hinese," in *Proceedings of the Sixth Workshop on {NLP} for Similar Languages, Varieties and Dialects*, 2019.

[13] G. de Francony, V. Guichard, P. Joshi, H. Afli, and A. Bouchekif, "Hierarchical Deep Learning for Arabic Dialect Identification," *Proc. Fourth Arab. Nat. Lang. Process. Work.*, pp. 249–253, 2019, doi: 10.18653/v1/w19-4631.

[14] Y. Fares *et al.*, "Arabic Dialect Identification with Deep Learning and Hybrid Frequency Based Features," 2019, doi: 10.18653/v1/w19-4626.

[15] M. Abbas, M. Lichouri, and A. A. Freihat, "ST MADAR 2019 Shared Task: Arabic Fine-Grained Dialect Identification," 2019, doi: 10.18653/v1/w19-4635.

[16] S. Wray, "Classification of closely related sub-dialects of Arabic using support-vector machines," *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 3671–3674, 2019.

[17] G. Lejeune, "Voting Classifier vs Deep learning method in Arabic Dialect Identification," *Proc. Fifth Arab. Nat. Lang. Process. Work.*, pp. 243–249, 2020.