Tinjauan Literatur : Identifikasi Dialek Dengan Deep Learning

by Rheza Daffa Pamungkas

Submission date: 24-Nov-2020 10:16PM (UTC+0700)
Submission ID: 1455062595
File name: Rheza_Paper_Automata_pake_dapus_-V5.pdf (244.38K)
Word count: 1874
Character count: 11639

Tinjauan Literatur : Identifikasi Dialek Dengan *Deep Learning*

***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. Didapati pada umumnya identifikasi dialek dilakukan dengan dua pendekatan yaitu pendekatan machine learning dan pendekatan deep learning. Tinjauan literatur ini memiliki tujuan untuk dapat digunakan sebagai referensi untuk pengembangan penelitian identifikasi dialek yang belum pernah dilakukan.**

***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 didaerah 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.

II. METODELOGI PENELITIAN

A. *Pertanyaan Penelitian*

Pertanyaan penelitian dapat menjadi langkah awal dari tinjauan literatur. Pada langkah ini pertanyaan dikemukakan dengan baik dan jelas guna membantu peneliti untuk mencapai tujuan yang diinginkan.Pertanyaan yang dikemukakan peneliti dalam mendasari penelitian ini dapat dilihat seperti berikut :

i. Bagaimana perkembangan tren untuk penelitian identifikasi dialek?
ii. Metode apa saja yang dapat digunakan dalam identifikasi dialek?
iii. Penelitian identifikasi dialek lebih banyak menggunakan dialek apa saja?
iv. Bagaimana saran yang dapat diberikan untuk mambantu penelitian identifikasi dialek di kemudian hari?

B. *Pengumpulan Data*

Data yang dikumpulkan yaitu 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 berupa 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 tabel 1 berikut.

Tabel 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 identifikasi dialek bahasa jawa memerlukan beberapa tahapan jika ingin mendapatkan hasil keluaran yang diharapkan.Tahapan-tahapan tersebut dapat dilihat sebagai berikut.

1) Pengumpulan data

Data yang dikumpulkan dapat beragam dan disesuaikan dengan kebutuhan yang sesuai dengan tujuan penelitian. Dapat dilihat pada Tabel 2 yang menunjukkan data apa saja yang digunakan pada penelitian-penelitian sebelumnya.

Tabel 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]	News 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]	*News*	*China dan Taiwan manadarin dengan menggunakan simplified Chinese dan traditional chinese*
[13]	*GDI and ADI*	*Arabic dialect dan German dialect*
[14]	Data menggunakan penelitian sebelumnya yaitu *MADAR (Multi Arabic Application and Resource)*	*Arabic Dialect*
[15]	Data diambil dari hasil wawancara dengan narasumber yang berbeda yang mewakili 5 dialect	*Kannada Dialect*
[16]	*MADAR Travel Domain dataset*	*Arabic Dialect*
[17]	Dataset Disediakan oleh organisasi MGB5 yang diambil dari *platform youtube*	*Arabic Dialect*

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 penelitiannya.Salah satu contoh pada penelitian Elaraby,et al[10], pada penelitian tersebut preprocessing yang dilakukan salah satunya adalah *tokenization and normalization* diperuntukkan data dengan spasi dan menghilangkan huruf *unicode* selain itu digunakna jua 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.

Fitur ekstraksi yang digunakan pada penelitian-penelitian sebelumnya dapat dilihat pada Tabel 3 berikut ini.

Tabel 3. METODE DANEKSTRAKSI FITUR YANG DIGUNAKAN

Referensi	Ekstraksi fitur
[6]	n-gram
[7]	
[8]	*Character level n-gram* *Word Alignment*
[9]	*Attention layer*
[10]	*Word embedding*
[11]	TF-IDF *Character level n-gram* *Word level n-gram*
[12]	
[13]	i-vector lexical
[14]	*Character* TF-IDF *Word* TF-IDF
[15]	
[16]	*n-gram*
[17]	

4) Pemodelan

. Pada tahapan selanjutnya yaiyu tahap pembangunan model yang akan digunakan untuk identifikasi dialek, Pembuatan model yang dilakukan pada penelitian yang lainnya dibuat sesuai dengan kebutuhan guna mencapai hasil output yang di inginkan.

Setelah pemodelan untuk identifikasi dialek dilakukan,maka langkah selanjutnya adalah Evaluasi atau pengecekan terhadap kinerja model yang sudah dibangun.Pengecekan ini dilihat dari besaran nilai output yang berupa F1 score atau Accuracy dari sebuah model. Tabel 4 dibawah ini menunjukkan tingkat hasil F1 *score* atau *accuracy* dari model-model yang telah dibuat oleh peneliti-penleiti sebelumnya.

Tabel 4. METODE DAN HASIL DARI PENELITAN SEBELUMNYA

Referensi	**Metode**	**Hasil**
[7]	LR SVM *language models*	F1 Score: 49.46% *and* 51.32%
[8]	CNN	*Accurcay:* 0.8307
[9]	SVM	*Accuracy :* Untuk berita 82%,90% dan 97% Untuk Wikipedia : 60% dan 77%
[10]	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*
[11]	SVM Multi-input CNN CNN-BiLSTM Binary Classification with CNN-BiLSTM	CNN with separate embedding (F1score : 0.5289), CNN-BiLSTM(F1score :0.4235) Binary CNN-BiLSTM (F1score :0.4339)
[12]	SVM	F1 Score : 0.6203
[13]	LR SVM MNB CNN RNN Bi-LSTM	F1 Score : 0.8530 0.8687
[14]	BiLSTM	*Accuracy :* i-vector: 0.577 lexical: 0.246
[15]	MNB Baseline LSTM	*F1 score* : 66.6%
[16]	SVM NN	*Accuracy :* SVM dan fitur MFCC 66% NN pada level kalimat dan level ucapan : 83.09% dan 92.71%
[17]	SVM MNB BNB LR SGD PA PC	*F1 Score :* 62%

[18]	CNN Transformer	*Accuracy :* 86.29%

KESIMPULAN

Berdasarkan kepada penelitian identifikasi dialek yang sebelumnya, tren penelitian identifikasi dialek dalam kurun waktu 2016 hingga 2020 adalah dapat menggunakan 2 pendekatan yaitu pendekatan *machine learning* seperti SVM, LR, dan MNB dan pendekatan *deep learning* seperti CNN, NN, RNN, LSTM, BiLSTM dan dapat dikombinasikan juga satu dengan yang lainnya. Metode tersbut juga diabntu dengan menggunakan beberapa ekstraksi fitur seperti *n-gram,* TF-IDF dan *attention layer*. Tetapi dilihat melalui tinjauan literatur ini,identifikasi dialek yang menggunakan Bahasa Indonesia terkhususnya dalam dialek Bahasa Jawa masih sedikit dilakukan. Maka untuk 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] 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.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[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] 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.

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

[15] 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.

[16] 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.

[17] W. Lin, M. Madhavi, R. K. Das, and H. Li, "Transformer-based Arabic Dialect Identification," -, 2020, [Online]. Available: http://arxiv.org/abs/2011.00699.

Tinjauan Literatur : Identifikasi Dialek Dengan Deep Learning

ORIGINALITY REPORT

5%	5%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	issuu.com Internet Source	1%
2	zombiedoc.com Internet Source	1%
3	www.garudacitizen.com Internet Source	1%
4	ejurnal.itenas.ac.id Internet Source	1%
5	www.delphion.com Internet Source	1%
6	ejournal3.undip.ac.id Internet Source	1%
7	ejournal.bsi.ac.id Internet Source	1%
8	www.scribd.com Internet Source	1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		