khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Performance of Methods in Identifying Similar Languages Based on String to Word Vector

Herry Sujaini
Department of Informatics
Universitas Tanjungpura
Pontianak
hs@untan.ac.id

Abstract-Indonesia has a large number of local languages that have cognate words, some of which have similarities among each other. Automatic identification within a family of languages faces problems, so it is necessary to learn the best performer of language identification methods in doing the task. This study made an effort to identification Indonesian local languages, which used String to Word Vector approach. A string vector refers to a collection of ordered words. In a string vector, a word is represented as an element or value, while the word becomes an attribute or feature in each numeric vector. Among Naïve Bayes, SMO, J48, and ZeroR classifiers, SMO is found to be the most accurate classifier with a level of accuracy at 95.7% for 10-fold cross-validation and 94.4% for 60%: 40%. The best tokenizer in this classification is Character N-Gram. All classifiers, except ZeroR shows increased accuracy when using Character N-Gram Tokenizer compared to Word Tokenizer. The best features of this system are the TriGram and FourGram Character. The TriGram is preferred because it requires smaller training data. The highest accuracy value in the combination experiment is 0.965 obtained at a combination of IDF = FALSE and WC = TRUE, regardless the conditions of the TF.

Keywords: identification of languages, regional languages, string to word vector

1. Introduction

Language identification functions to identify or recognize the language (or dialect) of a text. Language identification, whose task is to predict the natural language of a written text, is not one of the most challenging problems in computational linguistics but is very necessary for supporting the implementation of other computational linguistics such as machine translators. The accuracy of a Language Identification system is strongly influenced by the similarity of the languages that will be the target of predictions. This research will discuss the identification of very similar languages, namely Indonesian and Malay. Educational figures from Yogyakarta, Ki Hadjar Dewantara, revealed that the basic Indonesian language is the Malay language which is adjusted to its growth in Indonesian society [1]. This is what makes it sometimes difficult to distinguish between Indonesian and Malay. In this study, regional Malay languages, Malay Pontianak and Malay Sambas are used.

Malay Pontianak is one of the languages used by people in West Kalimantan Province. There is no accurate data that can show the number of speakers of languages spoken by Malay people in the city of Pontianak. Malay Pontianak language, in many of its vocabularies, is almost the same as Indonesian. This fact is because the Indonesian language originates and is rooted in Malay [2]. Malay Sambas or Sambas Dialect Malay (BMDS) is one of the regional languages in Indonesia. This language is spread throughout the Sambas Regency, West Kalimantan Province. Sambas Regency, with an area of 6,394.70 km2 or around 4.36% of the area of West Kalimantan Province, has a population of around 505,444 inhabitants [3].

Goutte [4] described the results of evaluations of language identification systems that are trained to recognize various languages. They investigate the progress made from one study to the next. They estimated the upper limit on the performance that can be achieved using voting and oracle plurality, and identify some very challenging sentences. The research uses many diverse languages, including Bosnian, Croatian, Serbian, Indonesian, Malaysian, Czech, and Slovak. The results of this study indicate that the learning curve can help to identify how the task is being studied and which language groups need to be further considered.

There is much research on language identification. One of these studies was presented by Zaidan and Callison-Burch [5]. The authors took resources taken from social media to create large data sets of informal Arabic that are rich in dialect content (more than 100,000 sentences) on three Arabic dialects: Levantine, Gulf, and Egypt. They marked the big data manually to dialect. The authors then use the collected labels to train and evaluate automatic classifiers for dialect identification and observe interesting linguistic aspects of the tasks and behaviour of annotators. By using an approach based on the assessment of language models, they developed a classification that significantly outperformed the baseline using large amounts of MSA data, even close to the level of accuracy shown by human annotators. Lu and Muhammed [6] in another study developed a system called LAHGA, which was positioned to classify HIV, the LEV dialect, the dialect, and the MAG dialect. The author identifies features manually by using these features from interesting devices using Tweets as a dataset for the training and testing process. They use three different classifications, namely the Naïve Bayes classification, the Logistic Regression classification, and the Support Vector Machine classification. During the manual testing process, they eliminate all noise and choose 90 tweets, 30 from each dialect, whereas, in 10-fold cross-validation, there are no human interventions. LAHGA's performance showed 90% in manual tests and 75% in cross-validation.

Other researchers conducted experiments using sentence level approaches to classify whether the sentence was MSA or Egyptian dialect on the task of classifying Arabic dialects [7]. They based their research on a supervised approach using the Naïve Bayes classification. The authors present a supervised approach to the identification of Arabic dialects at the sentence level. This approach uses the features of the underlying system for identification of the token level of Arabic Egyptian Dialect in addition to the core and other meta-features. The method used by them to decide on the choice of sentence given is MSA or EDA. They vary the size of LM on the performance of their approach and study the impact of two types of preprocessing techniques. The approach they used yielded much better accuracy than the previous approach. Safitri [8] conducted a study on the identification of spoken languages with phonotactics in Minangkabau, Sundanese, and Javanese languages, concluding that the PRLM Method showed the highest accuracy using telephone identifiers trained for English and Russian with an average of 77.42% and 75.94%.

Some researchers who have written the results of their research on String To Word Vector include Jhao et al. [9] who proposed a word insertion model at the sub-word level and a word vector generalization method that allows the addition of pre-training word insertions with fixed size vocabularies to estimate "word embeddings" of words that are outside the vocabulary. Other studies found that the F-measure of rhetorical categorization performance in scientific articles can be improved by using word labeling and semantic word representation by Word2Vec [10].

This study has a specific specification that is the application of the String To Word Vector method to identify local languages that have similarities with Indonesian. String To Word Vector methods encode documents into string vectors, not numeric vectors. The traditional approach to text categorization usually requires document encoding into numerical vectors. The approach used is machine learning-based for text categorization, where string vectors are accepted as input vectors, not numeric vectors. As a result, it can improve the performance of text categorization [11].

This paper discusses the performance of the Language Identification method, specifically for languages that have similarities based on the String to Word Vector.

2. Method

The data in this study used sentences in the three languages tested, namely Indonesian, Malay Pontianak, and Malay Sambas. Each language consists of 1,000 sentences so that a total of 3,000 sentences is used. Sentences in Indonesian are taken from internet sources and translated into Malay Pontianak and Malay Sambas.

The research instrument or tool used for data begging is the Waikato Environment for Knowledge Analysis (WEKA). WEKA provides an implementation of learning algorithms that can be applied easily to data sets. WEKA also includes various tools for changing datasets, such as algorithms for discretization and sampling. We can process data, process it into learning schemes, and analyze the classifiers they produce and their performance. All algorithms take their input in the form of a single relational table that can be read from a file or generated by a database request. One way to use WEKA is to apply the learning method to the dataset and analyze the results to learn more about the data [12].

This study uses a set of classifications provided by WEKA [13] by measuring the performance of several classifications with the research steps carried out can be seen in Figure 1.

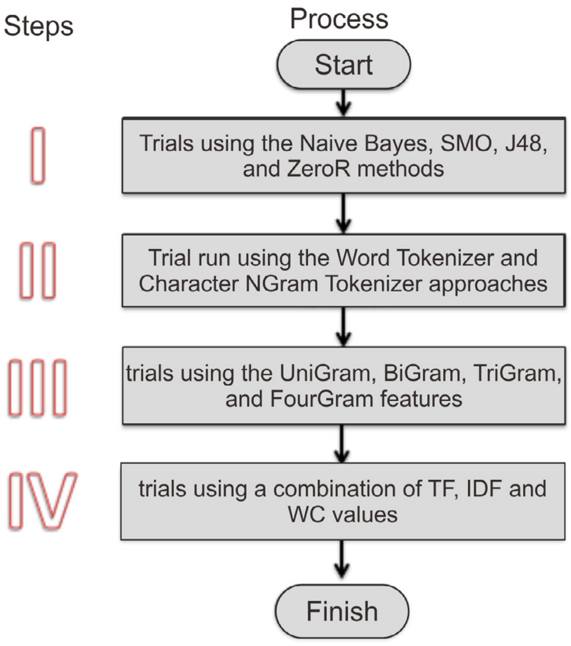

Steps
Process
Start
I
Trials using the Naive Bayes, SMO, J48, and ZeroR methods
II
Trial run using the Word Tokenizer and Character NGram Tokenizer approaches
III
trials using the UniGram, BiGram, TriGram, and FourGram features
IV
trials using a combination of TF, IDF and WC values
Finish

Figure 1. Research steps

In stage I, the experiment used 4 types of classification methods, namely Naïve Bayes, SMO, J48, and ZeroR. Each uses 10-fold cross-validation and 60%: 40% training data: test. In stage II, the experiment used 4 types of classification methods, namely Naïve Bayes, SMO, J48, and ZeroR, using 60%: 40% of training data: test. Each uses Word Tokenizer and Character NGram Tokenizer. In stage III, the experiment used 3000 sentences using 4 Character NGram features, namely UniGram, BiGram, TriGram, and FourGram, using the best classification algorithm from the results of step one and two experiments. Each using 10-fold cross-validation and 60%: 40% training data: test. In stage IV, the experiment uses a combination of different TF, IDF, and WC values using the best classification algorithm from experimental results 1 and 2, using the best Character NGram feature based on the experimental results in step three.

SVM (Support Vector Machines) works to find the hypothesis that reducing the boundary between correct errors in h will make it in the test data that is not visible, and errors in the training data. Sequential Minimal Optimization (SMO) is an implementation of the SVM classification of WEKA tools. SMO was developed for numerical prediction and data classification by building N-dimensions by optimally separating data into two categories [14]-[15]. SVM achieves the best performance in text classification tasks because SVM's ability to eliminate the need for feature selection means that SVM eliminates the high dimensional feature space that results from frequent occurrences of words in the text. Besides, SVM automatically finds proper parameter settings.

Naïve Bayes is one of the statistical classifiers, which can predict the probability of class membership of tuple data under the calculation of the probability of going into a particular class. The classifier discovered by Thomas Bayes in the 18th century is based on the Bayes theorem. In a comparative classification research report, a simple bayesian or commonly known as the Naïve Bayes classifier, shows high accuracy and speed when used in large databases [16].

J48 is one of the classifiers in data mining and part of a simple C4.5 decision tree. C4.5 builds a decision tree based on a set of labeled data inputs. A decision tree is a prediction model that uses tree structure or hierarchical structure. The decision tree has a concept in turning data into trees and decision rules [17].

ZeroR is the simplest classification method that depends on the target and ignores all predictors. ZeroR only predicts the majority category (class). Although there is no predictability in ZeroR, it is useful to determine baseline performance as a benchmark for other classification methods. Algorithm Build frequency tables for targets and choose their values most often. Contributors of Predictors Nothing can be said about the contributions of predictors to the model because ZeroR does not use one of them. The ZeroR evaluation model only predicts the majority class correctly. As mentioned earlier, ZeroR is only useful for determining baseline performance for other classification methods [18].

Term Frequency (TF) represents the frequency of specific keywords. Based on the data in the table, several words are usually found more often in one dialect than another dialect. So the weight of TF is used to show the level of importance of words in the text of the sentence. Inverse Document Frequency (IDF) scales how often a word appears in different sentence text (more than one dialect), which means words that appear in many dialects that cannot be used as features [19].

3. Results and Discussion

The data used are sentences in three languages, namely Indonesian, Malay Pontianak, and Malay Sambas. Each language consists of 1,000 sentences, so the total sentences used are 3,000 sentences, as in Table 1. The length of sentences used in this study ranged from 1-30 words, with an average of 18 words. The number of attributes (tokens) used is 4,349 tokens. Figure 2 shows the distribution of the number of words in the sentences used.

Table 1. Number of sentences used

Language	Sentence
Indonesian	1.000
Malay Pontianak	1.000
Malay Sambas	1.000

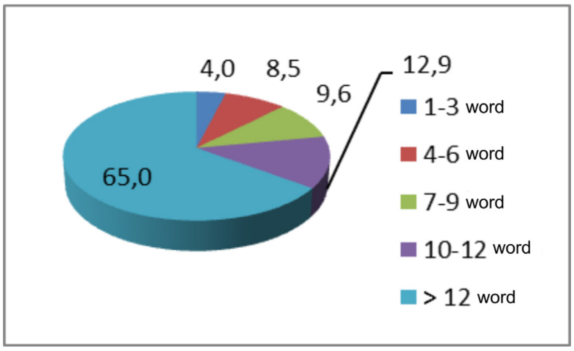
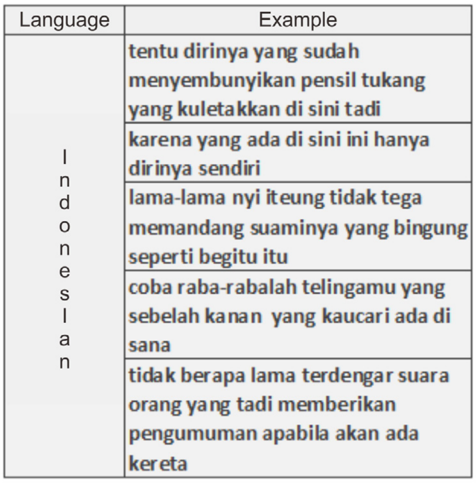


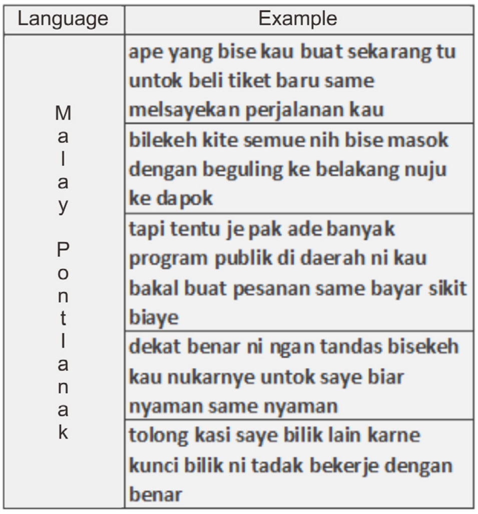
Figure 2. Distribution of the number of words in a sentence

Similarities between languages are characterized by words in sentences in a language. Similarities between Indonesian, Malay Pontianak, and Malay Sambas can be seen in the example of a few sentences in Figure 3-5. From these examples, it can be seen that several regional words are the same as the Indonesian language, for example, you, me, right, of course, an instrument, etc..



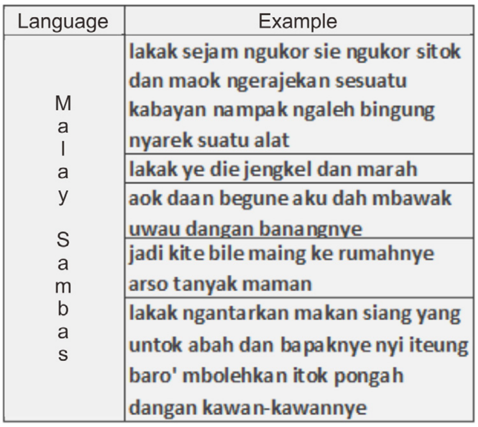
Language	Example
Indonesian	tentu dirinya yang sudah menyembunyikan pensil tukang yang kuletakkan di sini tadi
	karena yang ada di sini ini hanya dirinya sendiri
	lama-lama nyi iteung tidak tega memandang suaminya yang bingung seperti begitu itu
	coba raba-rabalah telingamu yang sebelah kanan yang kaucari ada di sana
	tidak berapa lama terdengar suara orang yang tadi memberikan pengumuman apabila akan ada kereta

Figure 3. Examples of Indonesian sentences



Language	Example
Malay Pontianak	ape yang bise kau buat sekarang tu untok beli tiket baru same melsayekan perjalanan kau
	bilekeh kite semue nih bise masok dengan beguling ke belakang nuju ke dapok
	tapi tentu je pak ade banyak program publik di daerah ni kau bakal buat pesanan same bayar sikit biaye
	dekat benar ni ngan tandas bisekeh kau nukarnye untok saye biar nyaman same nyaman
	tolong kasi saye bilik lain karne kunci bilik ni tadak bekerje dengan benar

Figure 4. Example of Malay Pontianak sentence



Language	Example
Malay Sambas	lakak sejam ngukor sie ngukor sitok dan maok ngerajekan sesuatu kabayan nampak ngaleh bingung nyarek suatu alat
	lakak ye die jengkel dan marah
	aok daan begune aku dah mbawak uwau dangan banangnye
	jadi kite bile maing ke rumahnye arso tanyak maman
	lakak ngantarkan makan siang yang untok abah dan bapaknye nyi iteung baro' mbolehkan itok pongah dangan kawan-kawannye

Figure 5. Example of Malay Sambas sentence

Table 2 reports the results for various classifications that were tried using the StringToWordVector filter with WordTokenizer, which is one of the WEKA features for extracting words as a feature of sentence strings. From table 2, it appears that SMO is the best classifier with an accuracy rate of 95.7% for 10-fold cross-validation and 94.4% for 60%: 40% of test and training data. While the lowest accuracy is obtained on the use of ZeroR for both experimental methods. From the ZeroR baseline, using SMO can increase accuracy by (0.957-0.333) / 0.333 = 187%.

Table 2. Classifier accuracy with different training methods

Classifier	10-fold cross-validation	60% : 40%
NaïveBayes	0,923	0,921
SMO	0,957	0,944
J48	0,836	0,812
ZeroR	0,333	0,325

The results of the classifier using Character NGram Tokenizer with Min = 3 and Max = 3 can be seen in Table 3. The Word Tokenizer method is a method for separating a series of words into tokens in the form of words or punctuation. The results of using this method show that the SMO classifier has a higher yield than the other classifier. Ngram Word Tokenizer has a function similar to Word Tokenizer. The difference lies in the function to enter the word order with the maximum and the minimum number of words, while Character NGram Tokenizer counts the combination of first, second, and so on, in sentence strings. The results of using this method show that the SMO classifier has a higher yield than the other classifier too.

Table 3. Classifier accuracy with word tokenizer and NGram tokenizer characters

Classifier	Word Tokenizer	Character NGram Tokenizer (3-gram)
NaïveBayes	0,921	0,931
SMO	0,944	0,965
J48	0,812	0,915
ZeroR	0,325	0,325

60% Experiment: 40% of this test and training data shows that all classifiers, except ZeroR have increased accuracy when using Character NGram Tokenizer compared to Word Tokenizer.

The first experiment to choose the best classification to identify Indonesian and Malay languages shows that the best classification of machine learning is the SMO algorithm. This study uses a WEKA StringToWordVector filter with Word Tokenizer that enters text into words between delimiters. But it was recommended to try n-Gram characters as units, not words as units. We used Character N GramTokenizer to divide strings into n-grams with maximum and minimum values. We set the Max value to 1, as well as the Min value on the model based on uni-gram; on the bigram model, we set Max to be 2, as well as the value of Min; on the tri-gram model, we set Max to be 3, as well as the Min value; in the 4-gram model, we set the Max value to 4, as well as the Min value. The results show that 4-gram models may not be appropriate because the training data is not large enough. Experiment using gram values that vary when evaluating by percentage of 60:40 from the training set and 10-fold cross-validation are shown in table 4.

Table 4. Feature accuracy with different training methods

Features	60% : 40%	10-fold cross-validation
Charater UniGram	0,809	0,828
Charater BiGram	0,935	0,943
Charater TriGram	0,965	0,966
Charater FourGram	0,965	0,967

The last experiment used a combination of different TF, IDF, and WC values using the best classification algorithm from experimental results 1 and 2, using the best Character NGram feature based on the results of previous experiments. The results show that the greatest accuracy value, 0.965, is obtained in combination TF = TRUE, IDF = FALSE dan WC = TRUE and TF = FALSE, IDF = FALSE and WC = TRUE. So it can be dreamed that a combination of values TF, IDF, and WC the best is IDF = FALSE dan WC = TRUE.

Table 5. Combination accuracy TF/IDF/WC

TF	IDF	WC	Precision
TRUE	TRUE	TRUE	0,960
TRUE	TRUE	FALSE	0,965
TRUE	FALSE	TRUE	0,969
TRUE	FALSE	FALSE	0,965
FALSE	TRUE	TRUE	0,960
FALSE	TRUE	FALSE	0,965
FALSE	FALSE	TRUE	0,969
FALSE	FALSE	FALSE	0,965

From the WEKA Confusion Matrix data testing 300 sentences each of 100 sentences that have been labeled as discussed, it is found that out of 100 Indonesian languages, two sentences are recognized as Malay Pontianak and two other sentences identified as Malay Sambas. Out of 100 Pontianak languages, one sentence is recognized as Indonesian and three sentences as Malay Sambas. While from 100 Malay Sambas languages, two sentences are recognized as Indonesian, and four sentences are recognized as Malay Pontianak.

4. Conclusion

This study classifies regional languages that are similar to Indonesian, namely Malay Pontianak and Malay Sambas, for the purpose of language identification. From Naïve Bayes, SMO, J48, and ZeroR classifiers, it was found that SMO was the most accurate classifier with an accuracy rate of 95.7% for 10-fold cross-validation and 94.4% for 60%: 40%. The best tokenizer in this classification is Character Ngram. All classifiers, except ZeroR have increased accuracy when using Character NGram Tokenizer compared to Word Tokenizer. The best features of this system are the TriGram and FourGram Character. TriGram is preferred because it requires smaller training data. The last experiment showed that the highest accuracy value, 0.965, was obtained in the combination of IDF = FALSE and WC = TRUE, regardless of the condition of TF.

References

[1] S. Sudaryanto, "Tiga Fase Perkembangan Bahasa Indonesia (1928—2009): Kajian Linguistik Historis", Aksis Jurnal Pendidikan Bahasa dan Sastra Indonesia , Vol. 2. No 1, 2018.

[2] E. Novianti, "Menilik Nasib Bahasa Melayu Pontianak". International Seminar Language Maintenance and Shiff. Pp. 70- 74. 2011.

[3] M.Z. Wiguna, "Tindak Tutur Bahasa Melayu Dialek Sambas di Kabupaten Sambas", Jurnal Pendidikan Bahasa, Vol. 5, No. 2, Desember 2016

[4] C. Goutte, S. Léger, S. Malmasi, and M. Zampieri, "Discriminating Similar Languages: Evaluations and Explorations", LREC, 2016.

[5] F. Omar, Zaidan and C.C. Burch. "Arabic dialect identification". Computational Linguistics, 40(1):171–202. 2014.

[6] M. Lu and M. Mohamed. "Lahga: Arabic dialect classifier". Report, December 13, 2011.

[7] H. Elfardy and M. Diab. "Sentence level dialect identification in arabic". In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, page 456-461. 2013.

[8] N.E. Safitri, A. Zahra, and M. Adriani, "Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages," Procedia Computer Science, vol. 81, pp. 182–187, 2016.

[9] J. Zhao, S. Mudgal, and Y. Liang, "Generalizing Word Embeddings using Bag of Subwords," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[10] G. H. Rachman, M. L. Khodra, and D. H. Widyantoro, "Word Embedding for Rhetorical Sentence Categorization on Scientific Articles," *Journal of ICT Research and Applications*, vol. 12, no. 2, p. 168, 2018.

[11] T.O. Ayodele. "Types of machine learning algorithms". 2010.

[12] M. Hall, E.Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. "The weka data mining software: An update". SIGKDD

Explorations, 11(1):10–18. 2009.

[13] I. H. Witten, and E. Frank. “Data mining: Practical machine learning tools and techniques with Java implementations”. San Francisco, CA: Morgan Kaufmann. 2016.

[14] T. Jo, “Representation of Texts into String Vectors for Text Categorization”. Journal of Computing Science and Engineering, 4(2), 110-127. 2010

[15] T. Joachims. “Text categorization with support vector machines: Learning with many relevant features”. 1998.

[16] F. Handayani and F. S. Pribadi, “Implementasi Algoritma Naïve Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110,” J. Tek. Elektro, 2015.

[17] S. Diwandari and N. A. Setiawan, “Perbandingan Algoritme J48 dan Nbtree untuk Klasifikasi Diagnosa Penyakit Pada Soybean,” Semin. Nas. Teknol. Inf. dan Komun., 2015.

[18] C. Nasa and S. Suman, “Evaluation of Different Classification Techniques for WEB Data,” Int. J. Comput. Appl., 2012.

[19] B.G. Gebre, M.Zampieri, P. Wittenburg, and T. Heskes. “Improving native language identification with tf-idf weighting”. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223. Association for Computational Linguistics. 2013.