Contents lists available at IdNNS
Indonesian Journal of Machine Learning and Intelligent Systems
Volume X ; Issue X
journal homepage: https://journal.idnns.org/index.php/ijmlis

Lexicon-Based Sentiment Analysis of Indonesian News Headlines on the Free Nutritious Meal (MBG) Program: TextBlob versus InSet

Panutan S. Sulendrakusuma[a,*]

[a]*School of Computer Science, Nusa Putra University, Sukabumi, Indonesia*

ARTICLE INFO

Keywords:
sentiment analysis
lexicon-based methods
Indonesian language
news headlines
TextBlob
InSet
query-term bias
public policy

ABSTRACT

Lexicon-based sentiment scoring is a common entry point in data-mining courses because it needs no training data. This paper applies a classroom pipeline built on Google News RSS, regular-expression preprocessing and TextBlob to 100 Indonesian headlines retrieved on 18 September 2026 with the query "Makan Bergizi Gratis", Indonesia's Free Nutritious Meal (MBG) program. Two Indonesian alternatives based on the InSet lexicon are added: a raw word-sum scorer and a refined scorer that masks query terms, removes stopwords and flips negated terms. All scorers are compared with a single-annotator reference set of 70 neutral, 19 negative and 11 positive headlines. TextBlob matched 0.2% of tokens and labelled 99 headlines neutral; its 71% accuracy is one point above the majority-class baseline ($\kappa = 0.05$). Raw InSet covered 31.7% of tokens but overpredicted polarity (macro-F1 0.33, $\kappa = 0.10$). The query word *gratis* (free) alone contributed +128 lexicon points, more than the corpus net score of +80. Masking query terms and stopwords reduced the net score to +15, turning an apparently favourable press into a balanced one, and raised negative recall from 0.47 to 0.58 without improving overall agreement. Off-the-shelf lexicons are therefore unreliable for Indonesian policy headlines, and query-term leakage should be controlled whenever sentiment is aggregated from keyword searches.

1. Introduction

News coverage is one of the few continuous signals on how a public program is being received and implemented. Automatic sentiment analysis promises to turn that coverage into a trend line, and lexicon-based methods are usually the first tool students and analysts reach for: they are transparent, fast and need no labelled data (Taboada et al., 2011; Liu, 2012).

This study starts from a data-mining lab notebook that retrieves headlines from the Google News RSS feed for a keyword, cleans them with regular expressions, scores them with TextBlob (Loria, 2026), and plots the label distribution and a word cloud. In the stored output of the original notebook (keyword "Pemerintahan Prabowo"), all ten headlines received a polarity of exactly zero. That result says little about the press. TextBlob relies on an English lexicon inherited from the Pattern library (De Smedt and Daelemans, 2012), so almost no Indonesian word can move the score.

We change the keyword to *Makan Bergizi Gratis* (MBG), the national Free Nutritious Meal program administered by the National Nutrition Agency (Badan Gizi Nasional, BGN). MBG coverage is dense and contested, covering budget absorption, food-safety incidents and the option for schools to decline the program, which makes it a demanding test for any sentiment method. Three questions guide the work. How much of the headline vocabulary do TextBlob and an Indonesian lexicon actually cover? How well do their labels agree with a human reading of the same headlines? Does the search query itself bias the scores?

The paper makes four contributions:

1) a reproducible extension of the lab pipeline to the MBG keyword with the Indonesian InSet lexicon (Koto and Rahmaningtyas, 2017) as a comparator;
2) an evaluation against manually labelled reference data using macro-F1 and Cohen's κ rather than raw accuracy;
3) the identification and quantification of query-term leakage, in which the lexicon weight of the search keyword dominates aggregate sentiment;
4) practical recommendations for classroom and monitoring pipelines.

*Corresponding author
panutan.sakti@nusaputra.ac.id (P.S. Sulendrakusuma)
ORCID(s):

Section 2 reviews related work, Section 3 describes data and methods, Section 4 reports results, Section 5 discusses implications and threats to validity, and Section 6 concludes.

2. Related Work

2.1. Lexicon-Based Sentiment Analysis

Lexicon-based methods assign prior polarity weights to words and aggregate them, often with rules for intensifiers and negation (Taboada et al., 2011). VADER extended the approach with heuristics for social-media text (Hutto and Gilbert, 2014), while work on contextual polarity showed that a word's prior polarity frequently differs from its polarity in context (Wilson et al., 2005). The main weakness of the family is domain dependence: a lexicon built for one genre transfers poorly to another (Liu, 2012; Pang and Lee, 2008).

2.2. TextBlob

TextBlob wraps the pattern-based analyser of Pattern, whose English lexicon annotates words with polarity in $[-1, 1]$ and subjectivity in $[0, 1]$ (Loria, 2026; De Smedt and Daelemans, 2012). In the version used here (0.20.1) the lexicon holds 1,528 distinct word forms, and 2,643 of its 2,912 part-of-speech-tagged entries (91%) are adjectives. Tutorials often apply it directly to non-English text, sometimes after machine translation.

2.3. Indonesian Resources

InSet is an Indonesian sentiment lexicon of 3,609 positive and 6,609 negative entries weighted from -5 to $+5$, developed for Indonesian microblog text (Koto and Rahmaningtyas, 2017). Supervised resources have since appeared, including the IndoNLU benchmark with its SmSA sentiment task (Wilie et al., 2020) and the IndoBERT language model (Koto et al., 2020). These require labelled data or fine-tuning and are outside the scope of the lab pipeline, but they define the natural next step.

2.4. Sentiment in News

News text differs from reviews and tweets. Journalists aim for neutral wording, headlines are short, and the positive or negative character of an item usually lies in the event reported rather than in evaluative vocabulary. Balahur et al. (2010) therefore identified three subtasks for news: defining the target, separating good and bad news content from the sentiment expressed on the target, and analysing explicitly marked opinion. Our annotation scheme follows that distinction.

3. Methodology

Fig. 1 summarises the pipeline. Stages A and B follow the lab notebook; stages C to E extend it.

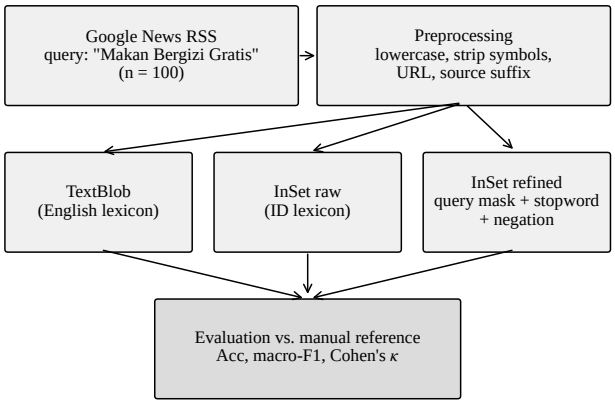


Figure 1: Analysis pipeline. Stages A and B follow the original lab notebook; the InSet scorers and the evaluation stage are additions.

3.1. Data Acquisition

Headlines were requested from the Google News RSS search endpoint with the notebook's URL template and the keyword replaced: `q=Makan%20Bergizi%20Gratis`, with Indonesian language and region parameters (`hl=id`, `gl=ID`). The notebook uses `feedparser` inside Google Colab. Because the execution sandbox used for this study blocked outbound requests to Google News, the feed was fetched on 18 September 2026 through a Chrome browser session and parsed for the same fields (`title`, `pubDate`). The feed returned 100 items. The notebook keeps only the first ten; all 100 were retained here to allow evaluation.

Publication dates range from 27 August to 18 September 2026, with 84 items published on 17 and 18 September. The items come from 60 distinct publishers; the most frequent one contributed 12 items, many of them syndicated versions of the same budget report. Two items are excerpts of Facebook posts rather than headlines and were kept as returned.

3.2. Preprocessing

The notebook's cleaning function lowercases the text, replaces every non-word character with a space, removes tokens beginning with `http`, and collapses whitespace. Two issues matter for Indonesian news. First, punctuation is removed before the URL pattern is applied, so a URL is already split into fragments when the URL rule runs and only the scheme is deleted. Titles in this corpus contain no URLs, so the order had no effect here, but it should be reversed. Second, Google News appends "- Publisher" to every title, so publisher names such as *kompas com* enter the analysed text.

TextBlob was run on the notebook's cleaned text without modification to preserve fidelity. For both InSet scorers, the trailing publisher suffix was stripped (repeatedly, since some titles carry two) before cleaning. After stripping, the corpus contains 1,099 tokens and

Table 1
Preprocessing of one headline.

Stage	Text
Raw title	Kepala BGN Akui Banyak Kasus Keracunan MBG, Ini Penyebabnya - Kompas.com
Notebook cleaning	kepala bgn akui banyak kasus keracunan mbg ini penyebabnya kompas com
Refined tokens	kepala bgn akui keracunan penyebabnya
Scores	TextBlob 0.0; InSet raw 0; InSet refined 0 (no lexicon match)

450 types, with a mean of 11.0 tokens per headline. Table 1 shows one headline at each stage.

3.3. Sentiment Scorers

TextBlob. TextBlob returns a polarity p between -1 and 1. Following the notebook, a headline is Positive if $p \geq 0.5$, Negative if $p \leq -0.5$ and Neutral otherwise. The sign rule ($p > 0$, $p < 0$) is also reported.

InSet raw. Let $M(h)$ be the lexicon entries found in headline h by greedy longest match over token n-grams, and $w(t)$ the weight of entry t, summed where a form appears in both the positive and negative lists. The headline score is

$$S(h) = \sum_{t \in M(h)} w(t) \tag{1}$$

and the label is

$$y(h) = \begin{cases} \text{Positive} & S(h) > k \\ \text{Negative} & S(h) < -k \\ \text{Neutral} & \text{otherwise} \end{cases} \tag{2}$$

with $k = 0$ unless stated. A variant that flips the sign of a term preceded by a negator was also computed.

InSet refined. Before matching, three filters are applied: (i) query terms $Q = \{$*makan, bergizi, gratis, mbg, program*$\}$ are masked; (ii) words in the 757-word Indonesian stopwords-iso list (Diaz, 2026) and numerals are removed; (iii) a negator (*tidak, tak, bukan, belum, jangan, tanpa*) flips the sign of the next retained term. None of these choices was tuned on the reference labels.

3.4. Reference Annotation

One annotator read every headline and labelled the event it reports about the program. *Negative* covers failures and harms: poisoning, spoiled food, kitchen suspensions, corruption probes, protests, budget discrepancies or cuts, and explicit criticism. *Positive* covers achievements, smooth delivery, beneficiary uptake and expressed benefits or support. *Neutral* covers budget figures, procedural and regulatory announcements, oversight measures, plans and hoax clarifications. The reference set contains 70 Neutral, 19 Negative and 11 Positive headlines.



Figure 2: Word cloud of MBG headlines after removing query terms, stopwords and publisher names.

3.5. Evaluation Metrics

We report accuracy, macro-averaged F1 over the three classes, and Cohen's κ (Cohen, 1960), computed with scikit-learn (Pedregosa et al., 2011):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where p_o is observed agreement and p_e the agreement expected by chance. Because 70% of the reference labels are Neutral, a classifier that always answers Neutral reaches 0.70 accuracy with $\kappa = 0$; macro-F1 and κ expose that case.

4. Results

4.1. Corpus Content

Excluding query terms, stopwords and publisher names, the most frequent terms are *bgn* (31), *anggaran* (budget, 27), *triliun* (15), *sekolah* (school, 13), *tolak* (refuse, 11), *bpom* (the food and drug authority, 10) and *keracunan* (poisoning, 9). Coverage in this window centres on the implementing agency, budget absorption, the rule allowing schools to opt out, and food-safety oversight (Fig. 2).

4.2. Label Distributions

Fig. 3 contrasts the reference distribution with each scorer. TextBlob labels 99 of 100 headlines Neutral. Only two corpus tokens exist in its lexicon: the loanword *top* (in "makin top", "even better") and *digital*. Its only negative score, -0.1, came from the English word *center* in the publisher name "Media Center Sembada". InSet behaves in the opposite way: it matches 31.7% of tokens and 92 of 100 headlines, and it labels about four in five headlines as polar.

4.3. Agreement with the Reference

Table 2 reports agreement. TextBlob's accuracy of 0.71 is one headline better than the majority baseline, and its κ of 0.05 confirms that the apparent performance comes from class imbalance. Raw InSet reaches 0.32 accuracy, macro-F1 0.33 and $\kappa = 0.10$;

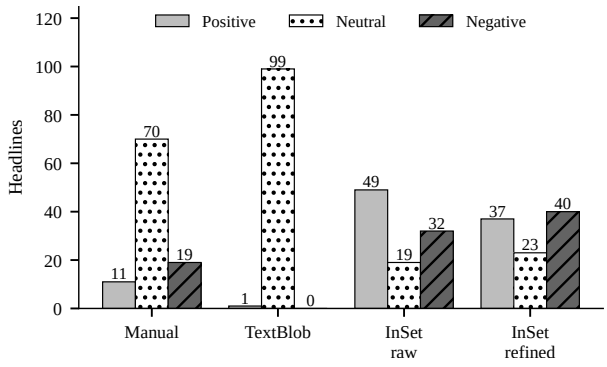


Figure 3: Label distribution of the manual reference and three scorers ($n = 100$).

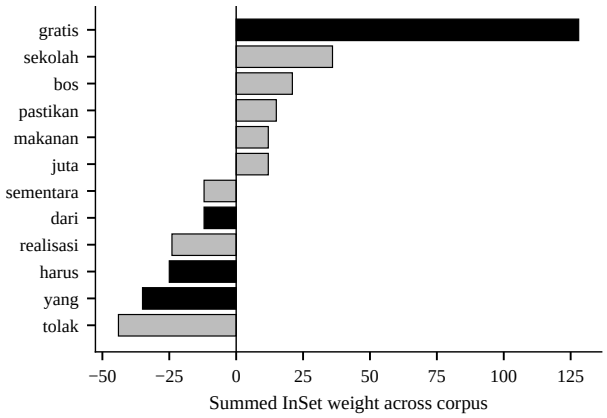


Figure 4: InSet entries with the largest summed weight under the raw scorer. Black bars mark the query word and function words removed by the refined scorer.

Table 2
Label counts and agreement with the reference ($n = 100$).

Method	Pos	Neu	Neg	Acc	$F1_m$	κ
Reference (manual)	11	70	19	–	–	–
Majority class	0	100	0	0.70	0.27	0.00
TextBlob ($\lvert p\rvert \geq 0.5$)	1	99	0	0.71	0.33	0.05
TextBlob (sign)	1	98	1	0.70	0.33	0.04
InSet raw	49	19	32	0.32	0.33	0.10
InSet raw + negation	50	17	33	0.31	0.32	0.10
InSet refined	37	23	40	0.33	0.32	0.07

$F1_m$: macro-averaged F1 over three classes.

Table 3
Confusion matrices of the InSet scorers (rows: reference; columns: predicted).

	InSet raw			InSet refined		
Reference	Pos	Neu	Neg	Pos	Neu	Neg
Positive (11)	10	0	1	6	3	2
Neutral (70)	35	13	22	27	16	27
Negative (19)	4	6	9	4	4	11

the refined scorer reaches 0.33, 0.32 and 0.07. Negation flipping changed only two labels. No scorer exceeds $\kappa = 0.15$, which falls in the "slight" band of the usual interpretation scale (Landis and Koch, 1977).

The confusion matrices in Table 3 show where the errors lie. Raw InSet labels 35 of the 70 neutral headlines Positive, which yields a Positive recall of 0.91 but a precision of only 0.20. The refined scorer shifts the balance: Negative recall rises from 0.47 to 0.58, but 27 neutral headlines are now labelled Negative.

4.4. Sources of Error

Fig. 4 lists the entries with the largest summed weight across the corpus under the raw scorer. Four error sources stand out; Table 4 gives examples.

Query-term leakage. The keyword *gratis* (free) carries a weight of +4 and occurs in 32 headlines, contributing +128 points. The net score of the whole corpus is only +80, so without this single query word the raw corpus would be net negative (−48). Masking query terms and stopwords brings the net score to +15. Any dashboard that averages lexicon scores over a keyword search inherits the polarity of the keyword itself.

Weighted function words. InSet assigns strong weights to high-frequency function words, including *yang* (which, −5, total −35), *harus* (must, −5, total −25) and *dari* (from, −3). Place names can also match by accident: *Agam*, a regency, scores +5.

Domain gaps. *Keracunan* (poisoning), the single most important negative event in this corpus, is absent from InSet. Of the eight reference-negative headlines containing it, the raw scorer labels two Negative and the refined scorer four. Policy words also carry misleading priors: *tolak* (refuse, −4) appears in 12 headlines, 11 of which report the neutral rule that schools may decline the program, and *realisasi* (realisation, −3) appears only in neutral budget reports.

Naive negation. In "angkanya tak masuk akal" (the figure makes no sense), flipping the next term turns *masuk* from −3 to +3, so the refined score becomes more positive for a clearly critical headline. Negation in Indonesian often scopes over an idiom rather than a single word.

4.5. Sensitivity to the Neutral Band

Since both InSet scorers over-assign polarity, we varied the neutral band k in Eq. (2). Fig. 5 shows that macro-F1 of the raw scorer rises from 0.33 at $k = 0$ to 0.45 at $k = 7$ ($\kappa = 0.15$), while the refined scorer peaks at 0.41 at $k = 1$. Because k was varied on the same 100 headlines used for evaluation, these figures are optimistic, and even the best stays below 0.5.

5. Discussion

5.1. Implications for the Lab Pipeline

The "all neutral" output of the original notebook should be read as a coverage failure, not as a finding

Table 4
Illustrative errors of the InSet scorers.

Headline (abridged), gloss and cause	Ref.	Raw	Refd.
Kepala BGN Akui Banyak Kasus Keracunan MBG. *BGN head admits many poisoning cases. Cause: keracunan not in lexicon.*	Neg	0	0
Pemkab Agam Perkuat Pengawasan MBG setelah Kasus Keracunan. *Agam tightens oversight after poisoning. Cause: place name Agam* $= +5$.	Neg	$+8$	$+5$
Sekolah Bisa Tolak MBG, BGN Ungkap Syarat yang Harus Dipenuhi. *Schools may opt out; BGN states conditions. Cause: yang, harus, tolak.*	Neu	-11	-1
Polemik Keracunan MBG ... Angkanya Tak Masuk Akal. *Poisoning estimate disputed as absurd. Cause: negation flips masuk.*	Neg	0	$+2$
Cegah Dikorupsi, ... Bikin MBG Makin Top. *Anti-graft steps make MBG "even better". Correct; TextBlob* $+0.5$ *via loanword top.*	Pos	$+9$	$+8$

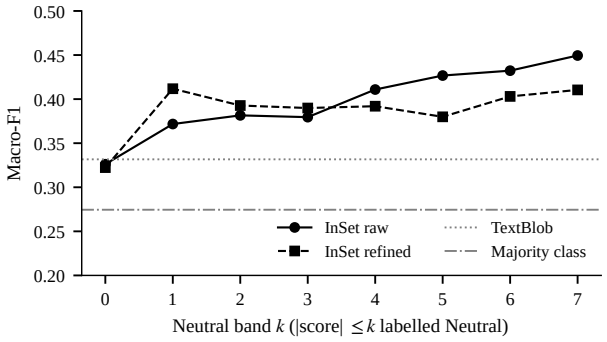


Figure 5: Macro-F1 as a function of the neutral band k ($|\text{score}| \leq k$ labelled Neutral). Exploratory: k is varied on the evaluation set itself.

about the press. A sentiment bar chart is only interpretable next to two numbers: the share of tokens the lexicon recognises and the score of a majority-class baseline. Both are cheap to compute. Reordering the cleaning steps and stripping publisher suffixes are small fixes that remove noise before any scoring.

5.2. Query Leakage in Aggregate Monitoring

The MBG case shows how a single lexicon entry can decide the headline conclusion of a monitoring report. With the raw scorer the press looks favourable; after masking the keyword it looks balanced; the human reference is predominantly neutral with twice as many negative as positive items. A program whose name contains a negative word would be biased the other way. Query terms should be masked by default whenever sentiment is aggregated from keyword searches, and the masked and unmasked results should both be reported.

5.3. Headlines Are Not Opinions

Most MBG headlines state facts. Their polarity lives in the event (a poisoning, a suspension, a corruption probe) rather than in adjectives, and a lexicon built from tweets captures little of it. Extending InSet with domain terms such as *keracunan*, *basi* (spoiled) and *disuspend* is a cheap stopgap but must be validated on held-out data. A more durable route is supervised classification, for example an IndoBERT model (Wilie et al., 2020; Koto et al., 2020) fine-tuned on the SmSA sentence-level sentiment data (Wilie et al., 2020) and adapted with in-domain news labels, or event-level annotation in the spirit of Balahur et al. (2010).

5.4. Threats to Validity

The reference labels come from a single annotator, so inter-annotator agreement could not be measured. The corpus is small and reflects one day's ranking of a personalised and opaque news service. Syndicated duplicates give some stories more weight than others. Only headlines were analysed. Finally, the labels describe the polarity of reported events for the program, not public opinion; none of the figures here should be read as a measure of how citizens view MBG.

6. Conclusion

Replacing the keyword in a standard lab pipeline with *Makan Bergizi Gratis* and adding an Indonesian lexicon reveals two failure modes. TextBlob does not read Indonesian and returns neutral labels whose accuracy is an artifact of class imbalance. InSet reads Indonesian but over-assigns polarity, weights function words heavily, misses the domain's key negative term, and lets the search keyword dominate the aggregate. Simple refinements change the overall picture but not the agreement with human reading. Future work will add a second annotator, collect a deduplicated corpus over several months, and compare the lexicon scorers with a fine-tuned IndoBERT classifier.

Data availability

The 100 headlines, reference labels, scores and analysis code are provided as supplementary files.

References

Balahur, A. et al. (2010). Sentiment analysis in the news. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

De Smedt, T. and Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, 13(66):2063–2067.

Diaz, G. (2026). stopwords-id: Indonesian stopwords. Stopwords ISO project, GitHub. https://github.com/stopwords-iso/stopwords-id. Accessed: 18 September 2026.

Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225, Ann Arbor, MI, USA.

Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 757–770, Barcelona, Spain (Online).

Koto, F. and Rahmaningtyas, G. Y. (2017). InSet lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 391–394, Singapore.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Number 16 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael, CA, USA.

Loria, S. (2026). TextBlob: Simplified text processing, version 0.20.1. https://textblob.readthedocs.io. Accessed: 18 September 2026.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Wilie, B. et al. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 843–857, Suzhou, China.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.