School of Computer Science Summer Research Scholarship AY2025/26

Project Title:	Stance Detection in Israeli Judicial Reform Discourse: An Annotated Hebrew Telegram Corpus and BERT-Based Benchmark
Supervisor Name:	Dr. Frank Glavin
Student Name:	Sade Gray

Stance Detection in Israeli Judicial Reform Discourse: An Annotated Hebrew Telegram Corpus and BERT-Based Benchmark

Sade Gray, Dr. Frank Glavin

University of Galway

Abstract. Israel's 2023 judicial reform produced one of the most consequential political controversies in the country's history, with Telegram a key venue for its contestation in Hebrew. We release an annotated Hebrew-language Telegram corpus of 2,116 messages labelled pro-reform, anti-reform, or neutral. Neutrality here denotes a genuinely balanced communicative position rather than a catch-all: posts offering no interpretable evidence of stance were set aside rather than absorbed into the neutral class, as a conventional favour/against/neither third category invites. The framework treats stance as a communicative or editorial act attributable to a post or channel rather than the private belief of every individual quoted within it, integrating stance-detection theory, Du Boisian accounts of stance, and media-gatekeeping research on selection, quotation, and framing; three annotators applied it with substantial-to-almost-perfect agreement ($\kappa = 0.80$). Benchmarking three pretrained Hebrew encoders, the two DictaBERT variants (184.3M parameters) outperform heBERT (109.5M) under five-fold cross-validation (macro F1 = 79.9% and 79.1% vs. 74.2%) and on an independent held-out set, indicating that model scale rather than parliamentary-domain pretraining drives performance. Across all models the dominant error is misclassifying neutral messages as pro-reform rather than confusing the two polarised stances with each other, supporting the framework's distinction between representational balance and communicative neutrality. The corpus, annotation codebook, and evaluation splits are released publicly under CC BY 4.0.

Keywords: Stance detection · Hebrew NLP · Telegram · Annotated corpus · Computational political science · BERT

1 Introduction

Israel's proposed judicial reform, advanced by the government formed in late 2022, generated one of the most consequential constitutional conflicts in the country's history [1]. The package included changes to the Judicial Selection Committee, limits on judicial review, changes to the reasonableness doctrine, and proposals affecting government legal advisers. Supporters framed it as a correction to excessive judicial intervention; opponents argued that it threatened judicial independence and the wider system of democratic constraints [2].

The debate is not reducible to a left–right split [3]: positions were frequently conditional or directed at particular components of the reform, and survey evidence from the period shows public support varying considerably across individual measures [4]. Panel evidence nonetheless identifies a stable pro- and anti-reform division, the camps differing in institutional trust and in how they weigh majority rule against minority rights [5]: the partition our labels operationalise.

Telegram has become a central venue for Israeli political communication, letting parties, activists, journalists, and protest movements distribute messages with limited editorial mediation; its channels can function as news outlets in their own right, with measurable patterns of source selection and coverage comparable to conventional media organisations [6]. Discourse on the reform therefore mixes explicit endorsements and condemnations with ambiguous, ironic, and context-dependent statements, which makes stance classification challenging [7, 8], especially in Hebrew, where political meaning often depends on morphology, framing, and knowledge of the controversies at issue. Pretrained Hebrew encoders such as AlephBERT [9], HeBERT [10], and the DictaBERT suite [11] are now available, and annotated Hebrew political-text resources have begun to appear alongside them: HebID labels social identities in Israeli politicians' Facebook posts [12], and a companion resource annotates delegitimization discourse across Knesset speech, Facebook, and news [13]. Neither annotates stance toward a policy target, and Hebrew is absent from the multilingual benchmarks on which cross-lingual stance transfer is evaluated [14]. Stance resources on Israel-related discourse do exist, but in English and Arabic and on the Israeli–Palestinian conflict rather than on domestic constitutional politics [15, 16]. To our knowledge this is therefore the first Hebrew-language stance-detection corpus, and the first on Israeli domestic politics.

This paper makes three contributions. First, we introduce an annotation framework that treats stance as a communicative or editorial act attributable to a post or channel, and that defines neutrality as a substantive position rather than a residual one. Established benchmarks adopt a favour/against/neither formulation [17] whose third category absorbs both genuinely balanced posts and posts offering no interpretable evidence of stance; we admit only the former, and set the latter aside. Prior work takes up each half separately: Caceres-Wright et al. separate irrelevant from neutral posts, with agreement rising once neutral cases are excluded [18], and discourse-analytic work treats apparent neutrality as itself a substantive position [19]. We combine the two and extend them to the channel as the unit of attribution. Second, we release the resulting corpus of 2,116 labelled messages, with the codebook and exact evaluation splits, under CC BY 4.0 (Sect. 5). Third, we benchmark three pretrained Hebrew encoders on it, with an error analysis locating the task's central difficulty at the boundary between balanced reporting and implicit endorsement rather than between the two opposing stances.

2 Corpus Construction

2.1 Collection and Keyword Filtering

The corpus was collected from publicly accessible Telegram channels through an authorised API client. Channels were identified by political relevance, activity during the period, and engagement with the reform, and were deliberately selected across a range of political orientations so as not to represent only one side of the debate. Messages were retrieved with all available metadata (Sect. 5).

The corpus was then filtered with twenty Hebrew keywords covering the terminology, institutions, and legal concepts of the debate: the reform itself (משפטית רפורמה, משפטית מהפכה), its specific measures (ההתגברות פסקת, "override clause"; הסבירות עילת, "reasonableness"), the institutions at issue (העליון המשפט בית, שופטים לבחירת ועדה), the normative vocabulary of the dispute (החוק שלטון, ובלמים איזונים, חוקתי משבר), and the protest movement (הרפורמה נגד הפגנות, הדמוקרטיה מחאת), with three orthographic variants of בג"ץ to accommodate variation in user-generated content. A message was retained if it contained at least one keyword, yielding 3,665 candidates; because keyword occurrence alone does not indicate a stance [7], no labels were assigned automatically. The full list ships with the dataset.

Preprocessing was deliberately conservative. URLs, mentions, and channel references were normalised or removed where they did not carry semantic content, whitespace was standardised, and extraction artifacts were cleaned. Stopword removal and stemming were not applied, since transformer encoders exploit context and function words, and because Hebrew is morphologically rich, a language in which sub-word segmentation materially affects performance [9, 10]; repeated punctuation was retained as potentially stance-relevant. Messages clearly unrelated to the political or institutional debate were then excluded, and the remainder submitted for manual annotation.

2.2 A Stance Framework for Channel-Level Political Discourse

Stance detection is conventionally defined as identifying the viewpoint expressed toward a specified target, typically as favour, against, or neither [7, 8]. Our scheme instead operationalises the construct as the stance *communicated* by a post or channel, rather than the private belief of every individual whose words appear within it. This matters on Telegram, where one message may combine original commentary, quotation, and forwarded material from several contributors. Four propositions define the scheme.

P1. Stance is an expressed position toward a target, explicit or inferable. Annotators first determine whether the reform is the target of a post, then whether the text communicates a supportive, opposing, or neutral position toward it; where none of the three can be established, the post is set aside rather than forced into a label. Following Burnham, who defines the task by whether a reader can reasonably infer support from a text [20], the scheme admits implicit as well as explicit cues: no sentence equivalent to "I support the reform" is required.

P2. Stance is a communicative act, not an aggregate of sentiment words. Following Du Bois's account of stance as a public act through which an actor evaluates an object, positions subjects, and aligns with or distances itself from others [21], stance may be conveyed through evaluation, source choice, alignment, distancing, framing, emphasis, and omission. It is separable from sentiment, which alone cannot recover it [22]: "The protesters were furious about the judicial reform" conveys negative affect without communicating the author's opposition, while "The government finally passed the reform, despite fierce opposition" has no valenced item directed at the reform yet frames it favourably. Annotators therefore weigh the target of evaluation and the post's positioning rather than counting valenced words.

P3. The channel is a legitimate editorial unit, and selection is evidence. A message such as "Levin: 'I love the reform"' contains two potentially relevant stances: Levin's own, and the stance the channel communicates by publishing it. Media-bias research distinguishes gatekeeping bias in the selection of stories and sources from coverage and statement bias [23], gatekeeping theory treats selection as structural rather than incidental [24], and Ameli and Molaei apply this at channel level to Telegram journalism [6]. The framework accordingly treats a quoted position as evidence of the channel's communicated stance when the channel selects, reproduces, or foregrounds it without distancing or counter-framing: not proof of private agreement, but observable communicative behaviour. For an owner-operated channel, repeated one-sided publication is evidence about the owner's position; for a multi-author or institutional channel, the construct is editorial stance.

P4. Neutrality is a substantive position, not a residual category. Two things are commonly mistaken for neutrality. The first is representational balance: "Levin says the reform protects democracy, while opponents say it undermines democracy" represents both sides but does not establish the channel's position, just as a one-sided post is not automatically an endorsement; balance is evidence contributing to the stance judgement rather than settling it. The second is absence of evidence. A genuinely non-partisan text is making a substantive communicative choice, stance-neutral formulation being a recognised rhetorical resource rather than a failure to take a position [19], whereas a text supplying too little to determine any position reflects a limit on what annotation can establish. Treating the latter as neutral, which a favour/against/neither third category invites, forces annotators to record a substantive judgement where none is warranted, a particular concern where opinion was itself differentiated across the reform's components [3, 4]. The scheme reserves the neutral label for the former and sets the latter aside (Sect. 2.3).

2.3 Annotation and Agreement

Three annotators independently applied the scheme to every screened message; with an odd number of coders, disagreements were resolved by majority decision without a separate tie-breaking procedure. Where an annotator was uncertain,

the protocol required a brief written justification, letting the others review the reasoning before a final label was assigned. Inter-annotator agreement was $\kappa = 0.80$ [25], at the boundary between substantial and almost perfect agreement [26] and a strong result given the interpretive difficulty of politically contested and frequently indirect stance expressions. Posts that discussed the reform but offered no interpretable evidence of a stance were set aside rather than labelled neutral (P4), yielding 2,116 stance-labelled messages: a 1,692-message development set and an independent 424-message held-out test set. Setting cases aside sits in tension with the argument that annotator disagreement is meaningful variation rather than noise and should not be discarded [27]. The criterion here is not disagreement: cases were removed where annotators independently converged on the judgement that a post carries no interpretable stance evidence, so what is excluded is absence of signal rather than contested signal, and stance disagreements were resolved by majority rather than set aside. Excluding such cases has been found to raise agreement substantially in comparable political stance annotation [18]. The complete codebook, including category definitions, the treatment of quotation and selective reporting, and worked examples, ships with the dataset (Sect. 5).

2.4 Corpus Composition

The development set contains 910 pro-reform messages (53.8%), 545 neutral (32.2%), and 237 anti-reform (14.0%).

The corpus is drawn from approximately thirty channels, but counts are highly concentrated: the largest channel contributes more than 1,300 of the 2,116 messages, and over a dozen contribute fewer than thirty each. Figure 1 shows this concentration (bar width) and each channel's stance composition (bar height). Channels differ sharply: some consist almost entirely of a single stance category, including at least one uniformly anti-reform channel, while others mix all three. This heterogeneity confirms that selection captured a range of political orientations, and shows why Sect. 2.2 treats the channel, rather than the corpus, as the unit for judgements about selection and framing.

Weekly volume is low and flat before the reform's introduction, rises sharply in early 2023, peaks that year, and continues at an elevated, episodic level through 2024–2026, with all three categories rising together during the peak and pro-reform messages most numerous throughout: the class imbalance is a stable feature of the corpus rather than an artefact of one window. Message length ranges from 16 to 4,148 characters, heavily right-skewed.

3 Benchmark Experiments

3.1 Setup

Three pretrained Hebrew encoders were fine-tuned for stance classification: DictaBERT [11] and Knesset-DictaBERT [28], two variants of the DictaBERT

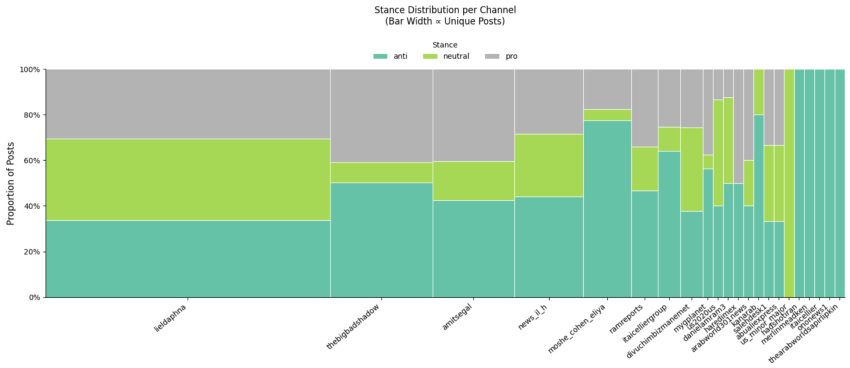


Fig. 1: Stance composition by source channel. Bar width is proportional to the channel's message count; bar height gives its proportion of anti-reform, neutral, and pro-reform messages.

architecture with roughly 184.3M parameters each, and heBERT [10], a smaller BERT-base model with roughly 109.5M. The Knesset variant is further pre-trained on Israeli parliamentary proceedings, giving a direct comparison between a general-domain Hebrew encoder and one adapted to institutional and political language.

The task is three-way classification over the 1,692-message development set, used for stratified five-fold cross-validation, with the 424-message held-out set (anti-reform, n = 60; neutral, n = 136; pro-reform, n = 228) untouched until the final generalisation check. Class skew was addressed by weighting cross-entropy inversely to class frequency. All models were trained with a single fixed random seed (42), so the dispersion reported below reflects variability across folds rather than across initialisations. Fine-tuning on datasets of this size is known to be sensitive to weight initialisation and data order [29], so our fold-level dispersion is a lower bound on total variability. We therefore treat only the between-model gap that survives bootstrap resampling as reliable, and read the remaining differences as directional.

3.2 Results

Pooling out-of-fold predictions across the five folds to obtain a single estimate over all 1,692 messages, DictaBERT achieved the highest cross-validated performance (accuracy = 83.2%, macro F1 = 79.9%), narrowly ahead of Knesset-DictaBERT and more clearly ahead of heBERT (Table 1). Bootstrap 95% confidence intervals for the two DictaBERT variants overlap substantially, so the difference between them is not reliable, whereas heBERT's interval does not overlap DictaBERT's and only marginally touches Knesset-DictaBERT's lower bound, indicating a robust gap between the larger architecture and the smaller model. Fold-to-fold variability in macro F1 was lowest for Knesset-DictaBERT (SD = 1.6 points), highest for heBERT (2.9). On the independent held-out set

the ranking of the two DictaBERT variants reverses, Knesset-DictaBERT obtaining the highest accuracy (78.8%) and macro F1 (74.7%). The three held-out macro-F1 scores span less than two points, and every model scores below its cross-validation estimate (by 6.0, 4.4, and 1.2 points respectively), consistent with the expected optimism of in-sample cross-validation.

Table 1: Aggregate and per-class performance, in percent. Cross-validation (CV) figures pool out-of-fold predictions over N = 1,692; held-out figures are for the independent N = 424 test set; per-class F1 is under CV, with class support in parentheses. Bootstrap 95% CIs for CV macro F1: 77.6–82.0, 76.8–81.3, and 71.9–76.7 respectively.

	CV		**Held-out**		**CV F1 by class**		
Model	**Acc.**	**Mac. F1**	**Acc.**	**Mac. F1**	**Anti** (237)	**Neutral** (545)	**Pro** (910)
DictaBERT	**83.2**	**79.9**	76.9	73.9	71.8	**79.4**	**88.3**
Knesset-DictaBERT	82.1	79.1	**78.8**	**74.7**	**72.8**	77.2	87.1
heBERT	78.5	74.2	76.7	73.0	64.3	73.7	84.7

All three models perform best on the majority pro-reform class and worst on the minority anti-reform class, mirroring the underlying distribution despite class-weighted training. Knesset-DictaBERT obtains the best anti-reform F1 (72.8%), narrowly ahead of DictaBERT (71.8%), despite trailing it on every aggregate cross-validation metric. The same ordering holds on the held-out set, where anti-reform remains lowest for every model (F1 = 63.6–64.6%) and pro-reform highest (82.6–84.8%).

3.3 Error Analysis

Across all three models the largest single source of error is misclassifying neutral messages as pro-reform (Table 2): 86 of 545 neutral messages under DictaBERT (15.8%), 107 under Knesset-DictaBERT (19.6%), and 113 under heBERT (20.7%), consistently more frequent than confusion in the opposite direction. Direct confusion between the two polarised classes is comparatively rare: at most 51 of 237 anti-reform messages (21.5%, heBERT) are predicted pro-reform, and at most 33 of 910 pro-reform messages (3.6%) anti-reform. The models therefore rarely mistake one polarised stance for the other; most residual error concerns the neutral–pro-reform boundary rather than a directional partisan bias.

4 Discussion

4.1 Scale Drives Performance; Domain Adaptation Helps the Minority Class

Model scale, rather than domain-specific pretraining, is the primary driver of performance here: both DictaBERT variants outperform heBERT throughout,

Table 2: Out-of-fold confusion matrices for DictaBERT (D), Knesset-DictaBERT (K), and heBERT (H). Rows are true labels, columns predicted; correct predictions in bold.

	Predicted anti			**Predicted neutral**			**Predicted pro**			
True label	**D**	**K**	**H**	**D**	**K**	**H**	**D**	**K**	**H**	**Total**
Anti-reform	**171**	**169**	**147**	32	32	39	34	36	51	237
Neutral	37	31	40	**422**	**407**	**392**	86	107	113	545
Pro-reform	31	27	33	64	70	88	**815**	**813**	**789**	910

and the bootstrap intervals in Table 1 place this gap beyond sampling variability. This cuts against the general expectation that continued in-domain pretraining improves downstream performance [30], but is consistent with Hebrew-specific evidence that advantages on morpho-syntactic tasks correlate poorly with performance on semantics-heavy ones [31]. A plausible mechanism is register: parliamentary proceedings share most of their vocabulary with general Hebrew, so adapting to them adds less than adapting to a lexically distinctive domain. Further pretraining on Knesset proceedings produces a smaller effect, but one consistent in direction across two independent views of the data: Knesset-DictaBERT trails DictaBERT on aggregate cross-validation metrics, yet achieves the best anti-reform F1 under cross-validation and the best accuracy and macro F1 on the held-out set. A plausible reading is that parliamentary-register pretraining sharpens sensitivity to the institutional and legal vocabulary that recurs disproportionately in anti-reform messages (judicial independence, checks and balances, constitutional crisis), while offering little additional purchase on the colloquial register that dominates the majority pro-reform class. Given the single seed this is a direction rather than a point estimate, but it supports a simple rule: choose the larger encoder for aggregate accuracy, the domain-adapted variant when minority-class F1 or out-of-sample robustness matters most.

4.2 The Neutral-Pro-Reform Boundary Is the Substantive Challenge

The error analysis links the classification results to the framework of Sect. 2.2. The dominant error is precisely the pattern P4 anticipates: a message that reproduces pro-reform framing without explicit distancing is linguistically close to one that endorses it, even where annotators judged the communicative stance to be balanced. The confusion is not specific to Hebrew or to social media. Roy et al. report neutral and agree as the most frequently confused pair in web-document stance detection, reasoning that a document discussing a claim without positioning itself resembles one that endorses it [32]. That the same asymmetry recurs here, under a scheme that removes indeterminate posts from the neutral class, indicates it is a property of the balanced-reporting register rather than an artefact of a particular label inventory.

The asymmetry is not a base-rate effect. Class imbalance would predict errors drifting toward the majority class, but neutral-to-anti-reform confusion is

less than half the neutral-to-pro-reform confusion for every model (Table 2). This points to a genuine asymmetry in how balanced reporting is linguistically constructed here: neutral messages more often adopt the procedural, descriptive register of pro-reform government communication than the overtly critical register of anti-reform commentary. The finding is substantive, about the language of political neutrality, rather than an artefact of the label distribution.

Because indeterminate posts were set aside, the neutral class evaluated here contains only messages judged to communicate a genuinely balanced position. That this comparatively clean class remained hardest to separate from pro-reform framing locates the difficulty in language and pragmatics rather than in a loose category.

4.3 Implications for Measuring Polarised Discourse

For work using a classifier of this kind to estimate the distribution of public stance toward a contested policy, the consistent direction of the dominant error has a practical consequence: aggregate estimates from raw predictions will overstate the prevalence of pro-reform messages at the expense of the neutral category rather than distributing error evenly. Corrected rather than raw category proportions should therefore be reported. The problem is long recognised in computational social science, where the quantity of interest is a distribution over categories rather than a set of individual labels, and classifier error must be modelled rather than assumed away [33, 34]; design-based estimators do this with exactly the ingredients released here, a fallible classifier plus a small expert-annotated gold set, and show that uncorrected estimates retain bias and invalid confidence intervals even at high surrogate accuracy [35]. The held-out set and released confusion matrices make such a correction directly available. A second implication concerns minority classes: with inverse-frequency loss weighting, anti-reform F1 reached the mid-sixties to low seventies from roughly 190 training examples per fold, but the gap to the majority class remains. Targeted collection is the natural complement, and the released channel metadata identifies which channels to target.

Three boundaries define the scope of these results, each testable with the released resource. The evaluation covers a single platform, though the framework and codebook are portable to X or Facebook; it covers encoder-based models fine-tuned for classification, leaving open a comparison with instruction-tuned generative models; and the corpus is a single cross-sectional sample, so testing whether classification difficulty shifts across phases of the controversy requires the temporal stratification the released timestamps make possible.

5 Dataset Description and Availability

To support reproducibility and reuse, we release the complete annotated corpus as a standalone public dataset, independently of the experiments above.

5.1 Data Format and Schema

The release comprises all 2,116 labelled messages, one row per Telegram message, in a delimited tabular file. Each row carries the stance label, the cleaned text, and the message's own Telegram metadata: `channel`, `message_id`, `author_id`, `date` (ISO 8601, UTC) and `edit_date`, `views` and `forwards` at collection, and `reply_to`, `is_reply`, and `is_forward`. The channel field supports the channel-level analyses of Sect. 2.2, the timestamps the temporal stratification above. Development and held-out sets are distributed separately, so the split can be reproduced exactly rather than approximated, together with the row indices defining the five folds, the fold-level predictions, the keyword list, the codebook, and the exact Hugging Face identifiers of the three benchmarked encoders (`dicta-il/dictabert`, `GiliGold/Knesset-DictaBERT`, `avichr/heBERT`).

5.2 Access, Licence, and Reuse

The dataset is released on Hugging Face as sade-gray/IsraJudReform-Stance under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. We release full message text rather than identifiers alone, judging that reproducibility and reuse value outweigh the residual contextual-integrity concerns discussed in the Ethics Statement, given that every message was public when collected. The channel and author identifiers are Telegram's own and disclose nothing beyond what was already publicly attached to the source messages. Users are asked to review the Ethics Statement before downstream use.

6 Conclusion and Future Work

This paper presented an annotated Hebrew-language Telegram corpus on Israel's judicial reform, with a stance-annotation framework that treats stance as a communicative or editorial act attributable to a post or channel and reserves the neutral label for genuinely balanced communication. Fine-tuning three Hebrew encoders showed that the larger DictaBERT-family models outperform heBERT, that parliamentary-domain pretraining yields a targeted rather than general benefit, and that the dominant error is confusion of neutral with pro-reform messages rather than of the two opposing stances, supporting the framework's distinction between representational balance and communicative neutrality.

Next steps are to rebalance the anti-reform class through targeted collection from the channels in Fig. 1, to replicate the scheme across platforms and controversies, and to compare fine-tuned encoders with instruction-tuned generative models alongside feature-attribution methods testing whether models rely on the cues annotators were instructed to use. Beyond the empirical case, the dataset, framework, and error analysis offer a template for studying how multi-sided political controversies are expressed, contested, and computationally classified in informal digital discourse.

Disclosure of Interests. The authors have no competing interests to declare.

Ethics Statement

This study used only publicly accessible Telegram channels, retrieved via an API client operating within Telegram's terms of service; no private channels, groups, or direct messages were accessed, and no attempt was made to de-anonymise operators. Public availability does not imply consent to research use [36]; the dataset is released for studying aggregate discourse patterns and benchmarking NLP methods, not for identifying or targeting individuals. Keyword- and channel-based sampling may not be representative. Stance classifiers built on this data carry dual-use risk, as they could be repurposed for surveillance based on inferred political views [37]; we do not endorse such use and urge downstream users to weigh it before deployment. No human subjects were recruited; annotators were research team members, and no personal data about them is reported.

References

1. Navot, S.: An overview of Israel's 'judicial overhaul': small parts of a big populist picture. Isr. Law Rev. **56**(3), 482–501 (2023)
2. Roznai, Y., Dixon, R., Landau, D.E.: Judicial reform or abusive constitutionalism in Israel. Isr. Law Rev. **56**(3), 292–304 (2023)
3. Porat, I.: Political polarisation and the constitutional crisis in Israel. Isr. Law Rev. **56**(3), 369–384 (2023)
4. Hermann, T., et al.: The Israeli Democracy Index 2023. Israel Democracy Institute, Jerusalem (2024)
5. Simunovic, D., Dorfman, A., Katzir, M.: Exploring motivated reasoning in polarization over the unfolding 2023 judicial reform in Israel. Commun. Psychol. **2**, 43 (2024)
6. Ameli, S.R., Molaei, H.: Election journalism: investigating media bias on Telegram during the 2017 Iranian presidential election. Digit. Journal. **8**(8), 975–991 (2020)
7. Küçük, D., Can, F.: Stance detection: a survey. ACM Comput. Surv. **53**(1), 1–37 (2020)
8. AlDayel, A., Magdy, W.: Stance detection on social media: state of the art and trends. Inf. Process. Manag. **58**(4), 102597 (2021)
9. Seker, A., et al.: AlephBERT: language model pre-training and evaluation from sub-word to sentence level. In: Proceedings of ACL 2022 (Long Papers), pp. 46–56 (2022)
10. Chriqui, A., Yahav, I.: HeBERT and HebEMO: a Hebrew BERT model and a tool for polarity analysis and emotion recognition. INFORMS J. Data Sci. **1**(1), 81–95 (2022)
11. Shmidman, S., Shmidman, A., Koppel, M.: DictaBERT: a state-of-the-art BERT suite for modern Hebrew. arXiv:2308.16687 (2023)
12. Mor-Lan, G., et al.: HebID: detecting social identities in Hebrew-language political text. In: Findings of EMNLP 2025. ACL (2025)
13. Rivlin-Angert, N., Mor-Lan, G.: The enemy from within: political delegitimization discourse in Israeli political speech. In: Proceedings of EMNLP 2025, pp. 16617–16637. ACL (2025)

14. Hardalov, M., et al.: Few-shot cross-lingual stance detection with sentiment-based pre-training. In: Proceedings of AAAI 2022, vol. 36, pp. 10729–10737 (2022)
15. Zafeiropoulos, M., et al.: A social media analysis of discourse on the Israel–Palestine conflict on Telegram. arXiv:2608.21385 (2026)
16. StanceNakba shared task: actor and topic-aware stance detection in public discourse. In: Proceedings of Nakba-NLP 2026 (LREC-COLING). arXiv:2606.12068 (2026)
17. Mohammad, S., et al.: SemEval-2016 Task 6: detecting stance in tweets. In: Proceedings of SemEval-2016, pp. 31–41. ACL (2016)
18. Caceres-Wright, A.R., et al.: Explicit stance detection in the political domain: a new concept and associated dataset. In: SBP-BRiMS 2024. LNCS, vol. 14972, pp. 3–14. Springer (2024)
19. Sawaki, T.: On the function of stance-neutral formulations: apparent neutrality as a powerful stance constructing resource. J. Engl. Acad. Purp. **16**, 81–92 (2014)
20. Burnham, M.: Stance detection: a practical guide to classifying political beliefs in text. Polit. Sci. Res. Methods **13**(3), 611–628 (2025)
21. Du Bois, J.W.: The stance triangle. In: Englebretson, R. (ed.) Stancetaking in Discourse, pp. 139–182. John Benjamins, Amsterdam (2007)
22. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Trans. Internet Technol. **17**(3), 1–23 (2017)
23. D'Alessio, D., Allen, M.: Media bias in presidential elections: a meta-analysis. J. Commun. **50**(4), 133–156 (2000)
24. Shoemaker, P.J., Vos, T.P.: Gatekeeping Theory. Routledge, New York (2009)
25. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)
26. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
27. Plank, B.: The 'problem' of human label variation: on ground truth in data, modeling and evaluation. In: Proceedings of EMNLP 2022, pp. 10671–10682. ACL (2022)
28. Goldin, G., Wintner, S.: Knesset-DictaBERT: a Hebrew language model for parliamentary proceedings. arXiv:2407.20581 (2024)
29. Dodge, J., et al.: Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. arXiv:2002.06305 (2020)
30. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of ACL 2020, pp. 8342–8360. ACL (2020)
31. Cohen, A.D.N., et al.: HeQ: a large and diverse Hebrew reading comprehension benchmark. In: Findings of EMNLP 2023, pp. 13693–13705. ACL (2023)
32. Roy, A., et al.: Exploiting stance hierarchies for cost-sensitive stance detection of Web documents. J. Intell. Inf. Syst. **58**, 1–19 (2022)
33. Hopkins, D.J., King, G.: A method of automated nonparametric content analysis for social science. Am. J. Polit. Sci. **54**(1), 229–247 (2010)
34. Card, D., Smith, N.A.: The importance of calibration for estimating proportions from annotations. In: Proceedings of NAACL-HLT 2018, pp. 1636–1646. ACL (2018)
35. Egami, N., et al.: Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. In: NeurIPS, vol. 36 (2023)
36. franzke, a.s., Bechmann, A., Zimmer, M., Ess, C.M.: Internet Research: Ethical Guidelines 3.0. Association of Internet Researchers (2020)
37. boyd, d., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**(5), 662–679 (2012)