Topic Modeling + Stance Detection: Kenyan Parliament

Final Report

Sebastian Chacon, Daniel Huff, Eric Long

CS 5322, Natural Language Processing

Professor Lin

May 12th, 2025

Abstract

Contracted through the OIT Research Department at Southern Methodist University, our team has implemented several Natural Language Processing and Machine Learning techniques to connect trends between African Politician Backgrounds to their stances on given topics with politics (ie Finance, Unemployment Rate, Agriculture, Public Education, etc). Our Methodology starts with deriving a set of relationships between our given data being: Biographical Data (web scraped previously by Sebastian), 992 Hansard Text documents from Kenyan Senate, Voting Records (present in ~18% of documents), and 19 given topics from our client. With this information we want to answer the following questions: How does a politician's background affect their decisions on bills in the senate?, Does their association with any given organization affect their voting outcome?, and can a politician's background affect what topic they are for or against?. To answer these questions we followed this methodology: isolate our two way relationship of bills to topics -topics to bills and feed through a Topic Modeling algorithm, then repeat the process except through the one way relationship of speakers (politicians) to stances on topics and feed through a Stance Detection algorithm. After outputting both of these algorithms, we can then integrate the biographical data and deliver a full analysis of trending stances on topics and stances. While this end product was not produced in this report, the project will continue to be worked on and finished post-report.

Introduction

Dr. Cloward in the SMU Department of Political Science is in the process of writing a dissertation on the intersection of international relations and comparative politics, with particular attention to non-governmental organizations, international norms and transnational activism, international aid, Sub-Saharan African politics, and gender and politics. Part of this report requires the completion of a smaller scale project targeted towards the Kenyan Senate and then to be scaled at higher levels through all governments in Africa eventually. Through our testing, we ended with running four models: BERTOPIC for Topic Modeling, base class BERT trained on 2020 Election Tweets for Stance Detection, and a comparison of Logistic Regression and Random Forest for binary stance classification on given pieces of dialogue.

Methodology

This report picks up midway into the research project defined in which we start with the following materials: 992 Documents of Hansard Standardized Texts, 19 Political Topics that we are interested in (ie Agriculture, Education, Unemployment Rate, etc), a list of Speakers from the Kenyan Senate that are notable, and Biographical data on these Speakers (ie College, NCO Associations, etc). As shown in Figure 1, we started by defining our given materials and deriving their relationships. We were able to pull the following relationships: Bills and Topics have a two way relationship in that Bills are associated with many topics and each Topic is associated with many bills. Next we see an equal relationship in that every Text has Speakers that have texts. So they will be treated the same. Finally, a one way relationship of Speakers having stances on any given subject, bill, topic. After establishing these, we broke our project into two approaches: creating outputs for Topic Modeling, and outputting some Stance Detection Models. We took these approaches by implementing SBERT and BERTOPIC for the Topic Modeling, then using a

base class pre-trained (2020 Election Data) BERT and two of our own ML Models: Random Forest, and Logistic Regression.

Starting with Topic Modeling, Sebastian oversaw the implementation of the BERTOPIC model that would take the 992 Texts, perform Optical Character Recognition (OCR going forward) that extracts the text from PDFs, preprocessed our text, and use Semantic Embeddings courtesy of SBERT to use in the BERTOPIC. The preprocessing stage was particularly tough given the input. Starting with the OCR, we had to implement a double layer OCR algorithm due to failure on some levels randomly during runs. Using PLIMAGE it would pick up only portions of texts, scratches on paper would identify as different characters and would ruin our text. We ended up having major success with PyMuPDF that would extract and make all 992 documents usable. After implementing so, there was much time spent on the preprocessing of stop words and stemming. The PorterStemmer would over aggressively stem where words such as “committee” would be stemmed into “comite” which would take away the entire semantic meaning of the word and end up getting removed from our lang detect later through the process. We also had to implement a custom list of stop words for: days of the week, months of the year, politician names, and an 80% threshold for commonly used words. Doing so would allow us to have more accurate and representative topics. From there we used a Lang Detect from NLTK to remove all words that are not in the english dictionary. This serves two purposes, first is to remove any swahili. Although they speak that language often, it is not the official language of Kenya, so we decided to remove it. Second is to remove all misspelled words as they are uncommon enough in unique combinations we can remove this confounding variable as a whole. After preprocessing we feed our clean beautiful data into the SBERT to create topic embeddings and then into the BERTOPIC for topic modeling and labeling.

Before we had started the process for stance detection, we were unaware of any voting records being in the documents. Our plan was to train a model on UK Parliament Hansard documents, which have very accessible voting records. However, after investigating the Kenyan parliament documents, we found that some did have voting records, which is useful for training our models and evaluating their accuracy. Our first method for stance detection was to use a BERT based model specifically trained on stance detection for election data from the 2020 election. At first, this model was evaluating a lot of the text as neutral, which makes sense because parliamentary dialogue is much less concise and polarizing as a tweet can be. We had to force the model to make decisions if it was leaning a certain way to prevent this. After the model predicts each text block as for or against, it counts whether each speaker had more 'for' or 'against' to predict which way they voted. The model was then tested on the voting records provided at the end of the document.

In our second approach for stance detection, we trained two classifiers, Logistic Regression and Random Forest, on the text of 15 documents from the Senate.csv file that contained voting records. To do this, we first created the necessary training dataset by parsing out the voting records (senator names and whether they voted "Aye" or "No") from the chosen documents and then labeled each of the spoken texts with how that speaker voted in the end. We then performed standard text preprocessing techniques on these texts, including stop words removal, lemmatization, and the use of a TF-IDF vectorizer to attain TF-IDF embeddings (BERT was not used simply because the length of texts often exceeded the maximum token limit for BERT). We then used this preprocessed data to train the Logistic Regression and Random Forest classifiers, performing 5-fold cross-validation for each model, and scoring them primarily on their Kappa score.

Results

After running the Topic Modeling script, we receive 5 Portions of output shown in Figure 4 a-e. The first output is by document showing what topic it represents (clustered to), the second gives us by topic the representative words and documents and how many documents have been filtered into each topic, the third shows word probabilities by topic for all words embedded, fourth shows us the by document topic probability and boolean for representative document, and our fifth is generated by utilizing OpenAI API to give descriptions of each topic. Looking at Figure 4B we see the BERTOPIC successfully clusters topics together such as: Covid, Pandemic, Hospital, Sickness identifying key concepts from the documents. We can use these in results to connect back to the Stance Detection to create our final desired output of correlation to biographical data. We also see in Figure 4E that OpenAI is able to read and create accurate descriptions by topic of how the documents are connected.

For our first model, we found a very high success rate. The accuracy was typically between 90-100% for any given document that had voting records. An example of a given document's result can be seen in Figure 5. There were some limitations, though. As stated earlier, only 18% of documents had voting records, so we could not evaluate the model's accuracy on the large majority of the documents. In addition to this, the model could only make predictions for those who spoke, not those who voted. This was a major roadblock, as in most documents, the majority of voters do not speak. While this is not a major issue in the larger project, it did prevent some insightful information for our purposes for this portion of the project. Our findings, though, were generally promising for the future of this project.

For the second approach of stance detection, looking at Figure 6, we clearly see that the Logistic Regression model outperformed Random forest in all three metrics; and more importantly, it achieved a Kappa score of 0.546, which suggests moderate agreement between the

Logistic Regression model and the ground truth of the dataset. As such, while there is still much room for improvement in this aspect of stance detection, these results reflect a promising baseline for future work to be built on.

Conclusion

This report showcases moderate success in both the topic modeling and stance detection aspects of the work. While the end product was not fully realized, we have established solid baselines for two major parts of the project. Given that we now have the stances of some speakers, as well as the topics to which the documents belong to, the next step is to connect a speaker's stance with which document they spoke in. In other words, we want to link the document topics to the stances, and by integrating biographical data into this relationship as well, we aim to reveal how a politician's background and affiliations may influence their stance on specific issues, their voting behavior, and the topics they choose to support or oppose.

Figures

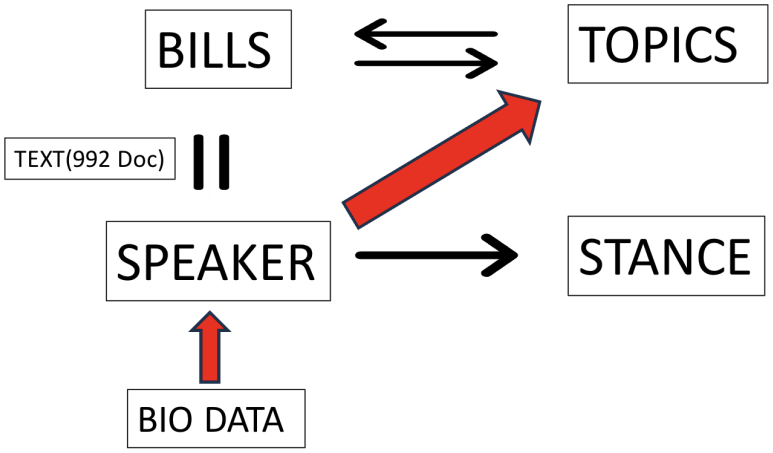
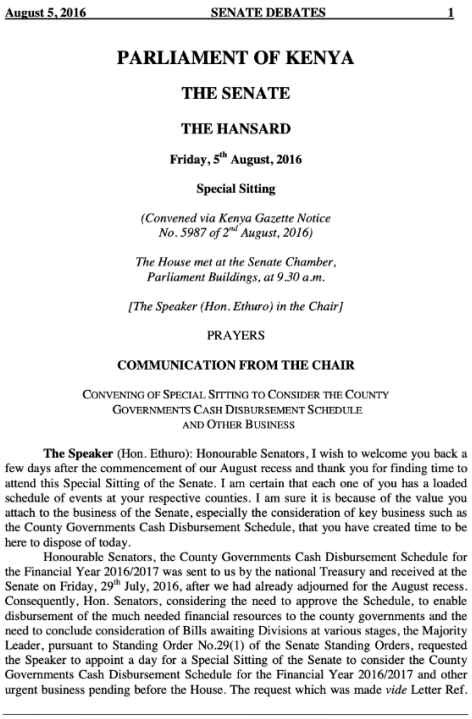


Figure 1:



August 5, 2016 SENATE DEBATES 1

PARLIAMENT OF KENYA

THE SENATE

THE HANSARD

Friday, 5th August, 2016

Special Sitting

(Convened via Kenya Gazette Notice No. 5987 of 2nd August, 2016)

The House met at the Senate Chamber, Parliament Buildings, at 9.30 a.m.

[The Speaker (Hon. Ethuro) in the Chair]

PRAYERS

COMMUNICATION FROM THE CHAIR

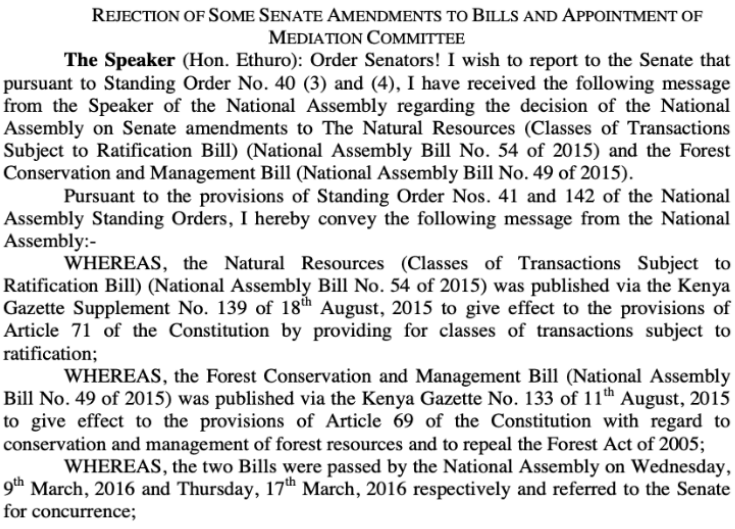
CONVENING OF SPECIAL SITTING TO CONSIDER THE COUNTY GOVERNMENTS CASH DISBURSEMENT SCHEDULE AND OTHER BUSINESS

The Speaker (Hon. Ethuro): Honourable Senators, I wish to welcome you back a few days after the commencement of our August recess and thank you for finding time to attend this Special Sitting of the Senate. I am certain that each one of you has a loaded schedule of events at your respective counties. I am sure it is because of the value you attach to the business of the Senate, especially the consideration of key business such as the County Governments Cash Disbursement Schedule, that you have created time to be here to dispose of today.

Honourable Senators, the County Governments Cash Disbursement Schedule for the Financial Year 2016/2017 was sent to us by the national Treasury and received at the Senate on Friday, 29th July, 2016, after we had already adjourned for the August recess. Consequently, Hon. Senators, considering the need to approve the Schedule, to enable disbursement of the much needed financial resources to the county governments and the need to conclude consideration of Bills awaiting Divisions at various stages, the Majority Leader, pursuant to Standing Order No.29(1) of the Senate Standing Orders, requested the Speaker to appoint a day for a Special Sitting of the Senate to consider the County Governments Cash Disbursement Schedule for the Financial Year 2016/2017 and other urgent business pending before the House. The request which was made *vide* Letter Ref.

Disclaimer: *The electronic version of the Senate Hansard Report is for information purposes only. A certified version of this Report can be obtained from the Hansard Editor, Senate*

Figure 2:



MESSAGE FROM THE NATIONAL ASSEMBLY

REJECTION OF SOME SENATE AMENDMENTS TO BILLS AND APPOINTMENT OF MEDIATION COMMITTEE

The Speaker (Hon. Ethuro): Order Senators! I wish to report to the Senate that pursuant to Standing Order No. 40 (3) and (4), I have received the following message from the Speaker of the National Assembly regarding the decision of the National Assembly on Senate amendments to The Natural Resources (Classes of Transactions Subject to Ratification Bill) (National Assembly Bill No. 54 of 2015) and the Forest Conservation and Management Bill (National Assembly Bill No. 49 of 2015).

Pursuant to the provisions of Standing Order Nos. 41 and 142 of the National Assembly Standing Orders, I hereby convey the following message from the National Assembly:-

WHEREAS, the Natural Resources (Classes of Transactions Subject to Ratification Bill) (National Assembly Bill No. 54 of 2015) was published via the Kenya Gazette Supplement No. 139 of 18th August, 2015 to give effect to the provisions of Article 71 of the Constitution by providing for classes of transactions subject to ratification;

WHEREAS, the Forest Conservation and Management Bill (National Assembly Bill No. 49 of 2015) was published via the Kenya Gazette No. 133 of 11th August, 2015 to give effect to the provisions of Article 69 of the Constitution with regard to conservation and management of forest resources and to repeal the Forest Act of 2005;

WHEREAS, the two Bills were passed by the National Assembly on Wednesday, 9th March, 2016 and Thursday, 17th March, 2016 respectively and referred to the Senate for concurrence;

Figure 3:

Figure 4:

a.

```
2020-06_Tuesday%2C%2016th%20June%2C%202020.pdf: Topic 34
2017-05_Thursday_26th_March_2015_-_Joint_Sitting_1.pdf: Topic 13
2017-05_Tuesday_3rd_June_2014.pdf: Topic 54
2023-02_Wednesday%2C%2015th%20February%2C%202023.pdf: Topic 49
2017-05_Thursday_17th_October_2013_1.pdf: Topic 147
2017-05_Tuesday_5th_May_2015.pdf: Topic 18
2023-01_Thursday%2C%2019th%20January%2C%202023.pdf: Topic -1
2021-10_Tuesday%2C%2012th%20October%2C%202021.pdf: Topic 159
2017-05_Tuesday_29th_September_2015.pdf: Topic 52
2021-02_Hansard%20-%20Tuesday%2C%2023rd%20February%2C%202021.pdf: Topic 160
2017-05_Wednesday_5th_November_2014.pdf: Topic 162
2022-10_Tuesday%2C%2025th%20October%2C%202022.pdf: Topic -1
2017-05_Tuesday_20th_February_2018.pdf: Topic 22
2017-05_Tuesday_8th_July_2014.pdf: Topic 241
2017-05_Tuesday_26th_April_2016.pdf: Topic 172
2019-11_Tuesday%2C%2019th%20November%2C%202019.pdf: Topic 25
2017-05_Hansard_Senate_17.04.2013P.pdf: Topic 96
2017-05_Thursday_6th_March_2014.pdf: Topic 70
2017-05_Thursday_22nd_March_2018.pdf: Topic 11
2022-02_Hansard%20-%20Tuesday%2C%2015th%20February%2C%202022.pdf: Topic 39
2017-05_Thursday_25th_September_2014.pdf: Topic 112
2017-05_Wednesday_20th_April_2016.pdf: Topic -1
2017-05_Thursday_26th_June_2014.pdf: Topic 116
2017-05_Tuesday_1st_July_2014.pdf: Topic 33
2021-11_Hansard%20-%20Tuesday%2C%2016th%20November%2C%202021.pdf: Topic 170
2017-05_Tuesday_26th_September_2017.pdf: Topic -1
2017-05_Wednesday_15th_February_2017.pdf: Topic 259
2017-05_Thursday_7th_August_2014.pdf: Topic 258
2017-05_Wednesday_4th_November_2015.pdf: Topic 18
```

b.

	Topic	Count	Name	Representation	Representative_Docs
0	-1	170	-1_substituting_therefor_immediately_oil	[substituting, therefor, immediately, oil, all...	[via gazette prayer auspicious outside led exc...
1	0	21	0_session_covid_debt_agenda	[session, covid, debt, agenda, recess, approva...	[prayer recess recess calendar month reflect r...
2	1	13	1_examination_corona_fish_coconut	[examination, corona, fish, coconut, tribunal,...	[visiting delegation acknowledge gallery visit...
3	2	11	2_milk_assumption_elect_clan	[milk, assumption, elect, clan, morocco, deleg...	[prayer visiting delegation pan african acknow...
4	3	9	3_wildlife_betting_gambling_manure	[wildlife, betting, gambling, manure, oversigh...	[prayer visiting delegation acknowledge presen...
..	...	...	...	...	...
264	263	2	263_figure_administrator_commerce_boy	[figure, administrator, commerce, boy, schedul...	[recall yesterday distinguished spoken era rec...
265	264	2	264_municipality_speech_flower_judiciary	[municipality, speech, flower, judiciary, urba...	[prayer loud commencement transact visiting de...
266	265	2	265_bank_child_primary_curriculum	[bank, child, primary, curriculum, treasury, b...	[violation registration bureau registration is...
267	266	2	266_participation_feedback_mining_principle	[participation, feedback, mining, principle, p...	[prayer hard copy kindly microphone checked un...
268	267	2	267_housing_payment_pesa_allocation	[housing, payment, pesa, allocation, omar, vat...	[prayer determination commencement serjeant ar...

c.

```
{-1: [('substituting', 0.003221806177838683), ('therefor', 0.0031588425602907958), ('immediately', 0.0029349229257059585), ('oil', 0.00248762937724045), ('allocation', 0.
0024861668785213 75), ('moi', 0.0024445748691420918), ('section', 0.0023159410247086017), ('research', 0.002286138377811464), ('voting', 0.0022565977514599016), ('sub', 0.
0022508579519958 9)], 0: [('session', 0.008389098921734929), ('covid', 0.004409741495339038), ('debt', 0.004360306270371547), ('agenda', 0.004263296268317571), ('recess',
0.003974146398224786), ('approval', 0.003851393636457654), ('legislative', 0.0037331315668423), ('pandemic', 0.003716512249985 4086), ('message', 0.003643367131898294), ('
scheme', 0.003449614883920802)], 1: [('examination', 0.008825458300455363), ('corona', 0.008236305954325023), ('fish', 0.008086837438048 37), ('coconut', 0.0080665435091452
86), ('tribunal', 0.007982789221824844), ('cheating', 0.007931707290487 33), ('attorney', 0.007862085114994574), ('virus', 0.006497102949386818), ('council', 0.00488443118
1610334), ('china', 0.004828725353539 04)], 2: [('milk', 0.007606570655798821), ('assumption', 0.006845266741256900 5), ('elect', 0.006176248615261552), ('clan', 0.00545447
3912834367), ('morocco', 0.005164426056076671), ('delegation', 0.004454183273621975), ('parliamentary', 0.003882746212862071 6), ('coalition', 0.003784336555340856), ('dep
ression', 0.0037725678824691 76), ('vending', 0.003736042101817522 5)], 3: [('wildlife', 0.013984597658955466), ('betting', 0.009681449641489738), ('gambling', 0.0084841147
2545418), ('manure', 0.007370088201588797), ('oversight', 0.006684650964430482), ('participation', 0.005666551318367 05), ('welcoming', 0.005370964222969229), ('justice',
0.005322160858346842), ('wild', 0.004973846725659544), ('collection', 0.004969754935266161)], 4: [('cancer', 0.01628924166221318), ('saccos', 0.01466662051555397), ('deaf
', 0.012441815285847736), ('kidney', 0.007659925181677 57), ('ken', 0.006650244277733665), ('sign', 0.006472824255169723), ('diabetes', 0.006424498637104792), ('language',
0.006103887946795922), ('census', 0.005854574179304082), ('extrajudicial', 0.005838502264528139)], 5: [('cancer', 0.026500533635139465), ('copy', 0.010262871153076473),
('editorial', 0.010066758157062 99), ('prevention', 0.010000071038457656), ('control', 0.007873481221600667), ('equipment', 0.007569549663812923), ('draft', 0.007514035948
126382), ('diagnosis', 0.006768250836910377), ('malaria', 0.006364683385212935), ('final', 0.005787514824584045)], 6: [('excellency', 0.04097871411889318), ('joint', 0.02
2896030688134426), ('eleventh', 0.020746021434770206), ('applause', 0.018897935853797 41), ('promote', 0.017059176976560 51), ('parliamentary', 0.015100291812892058), ('gro
wth', 0.014569057945933645), ('expand', 0.012658227649961 81), ('upstanding', 0.011951977905140672), ('spiritual', 0.011299678402202072)], 7: [('vocational', 0.01278402594
280919), ('delegation', 0.009168314604699692), ('training', 0.007989136904137389), ('sanitary', 0.007969655080035718), ('debt', 0.007943808278146 11), ('chute', 0.00719525
5123708814), ('estate', 0.006398243451714902), ('kindly', 0.00560550978558 3974), ('tabitha', 0.005567029354732254), ('fertilizer', 0.005479198730928421)], 8: [('potato',
0.014298752381263343), ('therefor', 0.011744989796202114), ('subclause', 0.011488242413197032), ('substituting', 0.011418102676394364), ('immediately', 0.009282719799256 4
67), ('council', 0.009012993402985144), ('printer', 0.008964107240781208), ('intelligence', 0.008125385859407 63), ('climate', 0.007730039601028 6125), ('paragraph', 0.0075
91676563980974)], 9: [('illness', 0.019015957469828 02), ('subsection', 0.014356192924657549), ('polytechnics', 0.013456061559986935), ('substituting', 0.012683876933569999
4), ('therefor', 0.012440650670425 93), ('ombudsman', 0.011344630946891747), ('supporter', 0.010195020129397938), ('paragraph', 0.009135655672309 93), ('facility', 0.008404
808198512866), ('unit', 0.007881154373393467)], 10: [('rung', 0.6648331395609336), ('bell', 0.5741142133168392), ('commencement', 0.21693025053438222), ('conference', 0.1
4484281452941153), ('determine', 0.09568570221947 32), ('expiration', 0.09303481989306182), ('adjourn', 0.08305085953131759), ('mandatory', 0.07188200595269705), ('ten', 0
.048170446778011 83), ('obviously', 0.03862754750930353)], 11: [('data', 0.032669451120983926), ('processor', 0.018737501601429875), ('energy', 0.010642343991808951), ('pe
```

d.

```
--- Document Info ---
                                                Document  Topic  ... Probability Representative_document
0    prayer message passage allocation message pass...    188  ...    1.000000                    True
1    migration television analogue digital platform...     -1  ...    0.000000                   False
2    prayer open tender fuel supply hereby federati...    115  ...    1.000000                    True
3    prayer indulgence intervene surely misappropri...     -1  ...    0.000000                   False
4    appreciation lady beyond zero campaign half ma...    105  ...    0.987008                    True
..                                                 ...    ...  ...         ...                     ...
987  prayer deferred lower irrigation project defer...     40  ...    0.914828                    True
988  prayer sessions joint parliamentary lay sessio...     83  ...    0.984512                    True
989  prayer compensation wildlife conflict petition...    140  ...    1.000000                    True
990  via gazette clarification nature significant b...     -1  ...    0.000000                   False
991  conference commencement confirm fate rapidly g...     -1  ...    0.000000                   False
```

e.

```
Topic 260:
Keywords: refugee, identity, potatoes, citizenship, camp, somali, insecurity, card, bag, registered
Description: The topic is about the process and challenges of Somali refugees, focusing on their identity and citizenship issues. It may also discuss their life in refuge
e camps, the importance of having registered identity cards, and basic needs like food (potatoes). The aspect of insecurity might refer to their uncertain and unstable li
ving conditions.

Topic 261:
Keywords: registrar, coalition, expulsion, kang, ata, nomination, ideology, register, indirect, definition
Description: The topic seems to be related to politics or political science. It could involve the process of registering for a political party, the formation of coalition
s, the expulsion of members, nominations for positions, and the ideologies or beliefs that guide these actions. It might also discuss indirect methods of influence or the
 definitions of these political terms.

Topic 262:
Keywords: textile, sugar, cane, clothes, industry, leather, china, company, sucrose, locally
Description: This topic is about the production and manufacturing industries, particularly focusing on textiles, clothing, and leather, as well as the sugar industry whic
h involves the processing of cane into sucrose. It also mentions that these industries are local to China and may involve specific companies.
```

Figure 5:

```
=== STANCE SUMMARY ===
Total FOR: 208
Total AGAINST: 6
Total NONE: 0

=== SPEAKER STANCES ===
Sen. Wakili Sigei: FOR
Sen. Mumma: FOR
Sen. Osotsi: FOR
Sen. Maanzo: FOR
Sen. Sifuna: FOR
Sen. Kisang: FOR
Sen. Kathuri: FOR
Sen. Gataya Mo Fire: FOR
Sen. Faki: FOR
Sen. Khalwale: FOR
Sen. Kinyua: FOR
Sen. Murango: FOR
Sen. Mandago: FOR
Sen. Okenyuri: FOR
Sen. Oburu: FOR
Sen. Methu: FOR
Sen. Veronica Maina: FOR
Sen. Omogeni: FOR
Sen. Wafula: FOR
Sen. Ogola: FOR
Sen. Munyi Mundigi: FOR
Sen. Mwaruma: FOR
Sen. Wamatinga: FOR
Sen. Abdul Haji: FOR
Sen. Tabitha Keroche: FOR

Evaluation Accuracy on voters: 20/20 = 100.00%
Total speakers in vote list: 48
Matched speakers with predictions: 20
Senators who voted but did not speak: {'dullo', 'chesang', 'tabitha keroche to second.', 'githuku', 'munyi', 'kavindu', 'thangw', 'lelegwe', 'methu to second.
', 'cheruiyot', 'mungatana', 'madzayo', 'seki', 'sifuna spoke off record', 'lomenen', 'kisang.', 'kavindu muthama', 'mungatana mgh', 'cheptumo', 'ali roba', '
```

Figure 6: