2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Linguistic Alignments

Detecting Similarities in Language Use in Written Communication

Lisa Kaati
Stockholm University
Stockholm, Sweden
Email: lisa.kaati@dsv.su.se

Amendra Shrestha
Mind Intelligence Lab
Uppsala, Sweden
Email: amendra@mindintelligencelab.com

Nazar Akrami
Uppsala University
Uppsala, Sweden
Email: nazar.akrami@psyk.uu.se

***Abstract*—Human language has many functions. Our communication on social media carries information about how we relate to ourselves and others, that is our identity, and we adjust our language to become more similar to our community - in the same way as we dress and style and act to show our commitment to the groups we belong to. Within a community, members adopt the community's language, and the common language becomes a unifying factor.**

In this paper, we explore the possibilities of identifying linguistic alignment - that individuals adjust their language to become more similar to their conversation partners in a community. We use machine learning to detect linguistic alignment to a number of different ideologies, communities, and subcultures. We use two different approaches: transfer learning with RoBERTa and traditional machine learning using Random forest and feature selection.

I. Introduction

In both face-to-face and online interactions, individuals tend to subconsciously and subtly adjust their language to become more similar to their conversation partners [7]. This linguistic accommodation not only enhances the efficiency of communication - it also creates a positive social identity, fostering a sense of sympathy and belonging - and indication identification with the group of individuals we are interacting with [20]. The motivation behind linguistic accommodation partly stems from a desire to express one's identification with a group, including a gain of social approval. The magnitude of identification would lead individuals to adapt more and embrace the linguistic style of the specific group. Linguistic innovation, such as creating new words or using existing words in novel ways, is recognized as a means to form subcultures [5] but also as a tool to express the group's unique identity and express group-specific matters more precisely. When an individual joins a community, their level of adaptation to the community's linguistic norms reflects their aspiration to fit in and the extent of their identity seeking. As new members, they can either adapt to existing community norms. Members who have been in the community for a long time can either adapt to the new norms or stick to their previous styles and be innovators who transform the community [6]. Research suggests that the more often people talk (or write) to each other, the more similar their speech becomes, meaning that users who communicate in the same community become linguistically closer over time [3]. Research also found women to accommodate the general structure of an online community to a greater extent than men [14].

In this work, we explore the possibilities of using machine learning to determine linguistic alignment with a set of online ideologies, communities, and subcultures. We have built classification models for determining linguistic alignment with seven different ideologies/subcultures/communities using data collected from online communication. The ideologies/subcultures/communities we consider are:

- ***Counter-jihad*** - a movement that considers Muslims living within Western boundaries a potential threat to Western society and culture.
- ***White supremacy*** - a belief that white people are superior to those of other races and thus should dominate them.
- ***Alt-right*** (alternative right) - an online phenomenon that can be described as a loosely connected far-right white nationalist movement.
- ***Animal rights*** - a movement promoting the idea that animals should be free to live without being used, exploited, or otherwise interfered with by humans.
- ***Environmentalism/environmental rights***- a movement aiming to protect natural resources and ecosystems.
- ***Incel*** - an online subculture consisting of men on incel forums that blame women and society for their lack of romantic success.
- ***Jihadist*** - an ideology promoted by terrorist groups such as the so-called Islamic State or Al-Qaida.

Previous work has focused on right-wing extremism and, in particular white supremacy. Yoder et al. [21] developed a classifier for detecting the language of white supremacist extremism. They used a Distil-BERT model fine-tuned with data from dedicated white supremacy forums such as Stormfront, Iron March, and the Daily Stormer. Alatawi et al. [1] developed two different classifiers for detecting white supremacist hate speech

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627594

TABLE I
THE DATA USED WHEN CREATING THE LINGUISTIC ALIGNMENT CLASSIFICATION MODELS.

Alignment	Data source	# texts
Animal rights	r/AnimalRights/	463
	Animal Liberation Front	1 911
	r/AnimalRebellion	22
Environment	r/environment/	14 756
	r/extinctrebellion/	476
	r/ClimateOffensive/	502
	Earth Liberation Fornt	211
Incel	Incels	2 613
	Blackpill Club	181
	Non-cucks United	81
	Looksmaxxing forum	168
	Looks theory	490
	Yourenotalone	65
Jihadist	Al-Risalah	83
	Dabiq	280
	Inspire	802
	Al-Hayat IS Report/News	459
	Rumiyha	128
	IS supporter blogs	10
Counter-jihad	Gates of Vienna	1 368
Alt-right	Daily Stormer	1 207
White supremacy	Stormfront	1 734
	VNN Forum	1 693
Normal population	Google blogs	300
	Boards	300
	NeoGAF	300
	Reddit	300

on Twitter. The first approach utilizes a bidirectional Long Short-Term Memory (BiLSTM) model, and the second one uses transfer learning with a BERT model. Siegel et al. [16] built a dictionary-based classifier that they used to estimate the levels of white supremacy communication on Twitter after the 2016 US election.

Further, several attempts have been made to detect jihadist propaganda or promoters of jihadist ideologies [2], [10], [18]. Most attempts have used Twitter data and various forms of machine learning to classify Twitter accounts as pro-IS or normal. However, concerns have been raised regarding the quality of the data and the data collection methods since the methods used are prone to sampling biases, and the datasets are not sufficiently filtered or validated [12]. Other work has examined methods for detecting incel communication [8].

II. LINGUISTIC ALIGNMENTS

The data (see Table I) we have used to train our classifiers is limited to certain groups, sources, or forums and does not cover an entire ideology or subculture.

- **White Supremacy (WS):** we used data from Stormfront and VNN Forum. Stormfront is presented as a community of "racial realists, idealists and white nationalists". The Vanguard National News Forum (VNN Forum) was launched in late 2001 as an uncensored forum for "white" people [9].
- **Counter-jihad:** we used data from Gates of Viennam, a website affiliated with the counter-jihad movement, featuring contributions from multiple writers. Gates of Vienna covers various aspects of the counter-jihad movement's historical evolution and offers information about European counter-jihad conferences.
- **Alt-right:** we used data from the Daily Stormer - one of the most notorious websites for the alt-right movement, established in 2013. The website gained attention when derogatory remarks about a woman who was killed in connection with the Unite the Right rally in Charlottesville during August 2017 was published.
- **Incel:** we have used data from six different incel forums: Incels, Blackpill club, Non-cucks united, Lookmaxxing forum, Looks theory, and Yournotalone. All these forums are dedicated meeting places for incels. Incels have developed their own characteristic language, their own areas of interest, and speculative theories that strengthen their members' cohesion and sense of belonging [13].
- **Animal rights:** we used data from Animal Liberation Front (ALF) and two subreddits. The ALF is an international group focused on animal rights with a website where recommendations for reading and information about actions are shared and commented on. The subreddits that we have used are r/AnimalRights/ and r/AnimalRebellion.
- **Environmentalism:** for environmentalism (environmental rights), we have used data from three different subreddits: r/environment, r/extinctrebellion, r/ClimateOffensive. We have also used data Earth Liberation Fronts (ELF) website. ELF is organisation that evokes direct action and revolutionary violence relying on a leaderless resistance model of operations [17].
- **Jihadist:** we have used the IS-produced magazines Dabiq and Rumiyah aimed specifically at the West [4]. We have also used the IS magazines Al-Hayat IS Report and Al-Hayat IS News and Al-Qaeda in the Arabian Peninsula's English language publication Inspire. Another source is the Al-Risalah an English-language propaganda magazine published by Jabhat al Nusra and extracts from IS supporter blogs.

III. METHOD

The data we have used for training each classifier is listed in Table I. For all forums and subreddits we collected data from users who have posted more than 20 and less than 10 000 posts in English. To determine the language of a post, the Python version of the library langdetect [15] was used. All posts from a user are merged into one text. For the magazines (Al-Risalah, Dabiq, Rumiyha, Inspire, Al-Hayat IS Report, and Al-Hayat IS News), we divided each magazine into articles or pages. The

TABLE II
THE DATA USED FOR THE POSITIVE AND THE NEGATIVE CLASS OF EACH LINGUISTIC ALIGNMENT MODEL.

Alignment	Positive class	Negative class
WS	Stormfront VNN Forum	Animal rights Environmentalism Jihadist Normal population
Alt-right	Daily Stormer	Animal rights Environmentalism Jihadist Normal population
Counter-jihad	Gates of Vienna	Animal rights Environmentalism Jihadist Normal population
Jihadist	Al-risalah Dabiq Inspire Al-Hayat IS Report Rumiyha Al-Hayat IS News Blogs	Animal rights Environmentalism White supremacy Alt-right Counter-jihad Normal population
Incel	Incels Blackpill Club Non-Cucks United Lookmaxxing forum Looks Theory Yourenotalone	Animal rights Environmentalism Jihadist Normal population
Environment	Earth Strike r/environment r/ClimateOffensive/ ELF website	White supremacy Alt-right Counter-jihad Incel Jihadist Normal population
Animal rights	ALF website r/AnimalRights/ r/AnimalRebellion	White supremacy Alt-right Counter-jihad Incel Jihadist Normal population

The data used as normal population is listed in Table I

normal population in Table I consists of 300 randomly selected users from three discussion forums and a set of blogs. Before training the classifiers, the data was cleaned. Each character was converted to lowercase.

For each linguistic alignment, term frequency-inverse document frequency (TF-IDF) was calculated, and the 1000 terms with the highest TF-IDF score were selected. The 500 most frequent word and bi-collocation words having a frequency of more than 50 were also extracted. All words (extracted using TF-IDF, most frequent, and bi-collocation) were manually annotated and are used as features for each linguistic alignment. While building the TF-IDF vocabulary features, words that appear in more than 20% of the documents and words that appear in less than 0.1% of the documents were removed. This process eliminates the most common words and words that seldom appear in the corpus. Table III shows some examples of features for each linguistic alignment.

TABLE III
EXAMPLES OF FEATURES FOR EACH LINGUISTIC ALIGNMENT.

Data	Example of features
WS	white, jews, (((, racial, population, truth
Alt-right	jews, soyboy, wanker, libtard, thots, whores
Counter-jihad	people, Muslim, western, Islam, democracy
Jihadist	jihad, messenger, soldiers, mujahidin, kufr
Animal rights	animal, cow, vegan, think, activist, slaughter
Env.	climate, water, power, carbon, money, global
Incel	foid, cucked, fakecel, truecel, stacy, girl

We trained two models: a RoBERTa model, and a Random forest model. Robustly Optimized BERT Pretraining Approach (RoBERTa) is a language model based on transformer architecture [19]. We utilized a pre-trained RoBERTa model made available from the transformers library Hugging face[1] and fine-tuned it with our datasets. Since most of the posts used for the experiment are longer text, the maximum sequence of tokens was fixed to 512 tokens. The experiment was done with five epochs, and the batch size was 8. During the training process, we chose the best-performing model measured by accuracy on the validation set. In the case with RoBERTa, Adam optimizer was used with a small learning rate of 5e-6.

To build the Random forest (RF) model, we used a bag of words model with manually selected features and the classification algorithm Random forest. When training the model, hyper-parameter tuning was done using grid search to estimate the optimal parameters of the classifier. The data was divided into 80% training and 20% test.

For each subculture/ideology, two different classes were created: a positive class and a negative class. The positive class contains data that represents the subculture/ideology, i.e., communication from a digital environment that is produced by members or promotors of the subculture/ideology. The data used for the negative class for each classifier is selected to represent subcultures/ideologies that it is very unlikely that an individual that belongs to the positive class is inspired by. For example, it is unlikely that an individual whose writing is aligned towards Alt-right also is aligned towards Animal rights and Environmentalism. However, an individual whose writing is aligned toward Alt-right might also be aligned toward White supremacy or Counter-jihad. The negative and the positive classes used to train each linguistic alignment classifier are presented in Table II. The normal population is described in Table I

[1]https://huggingface.co/docs/transformers/index

IV. Result

After training the Random forest model on 80% of the data, the model was tested on the resulting 20%. For the Random forest model - White supremacy, Alt-right, and Environmentalism had the highest F1-score.

When training the RoBERTa model, the data was divided into 70% train, 10% validation, and 20% test. The RoBERTa models had close to perfect F1-scores. The results are shown in Table IV. As seen, RoBERTa perform better than the Random forest.

TABLE IV
Results for the classification of linguistic alignment using Random forest and RoBERTa.

Model	Linguistic alignment	Precision	Recall	F1-score
Random Forest	Animal rights	0.96	0.93	0.94
	Environmental	0.97	0.98	0.97
	Incel	0.92	0.98	0.95
	Jihadist	0.98	0.94	0.96
	Counter-jihad	0.99	0.92	0.95
	Alt-right	1.00	0.93	0.97
	White supremacy	0.99	0.94	0.97
RoBERTa	Animal rights	0.99	1.00	0.99
	Environmental	0.99	1.00	1.00
	Incel	0.99	0.99	0.99
	Jihadist	0.99	1.00	0.99
	Counter-jihad	0.99	0.99	0.99
	Alt-right	0.99	1.00	0.99
	White supremacy	0.98	0.98	0.98

V. Testing in the wild

To test our linguistic alignment models in a more realistic scenario, we have tested the models on a set of texts that either are transcripts of speeches or written texts. The texts relate to various ideologies, communities, and subcultures we have trained our linguistic alignment classifiers to recognize. Table V shows the result of two different classifications where R is the RoBERTa model and RF is the Random forest model. When classification results are what we expected, they are in bold font.

The RoBERTa Counter-jihad model and the RoBERTa Alt-right model classified none of the text as Counter-jihad or Alt-right. The Random forest Counter-jihad model classified Brevik's text as Counter-jihad, and the Random forest Alt-right model classified the texts written by John Earnest, Brenton Tarrant, Peyton Gendron, and Dylan Roof as Alt-right (above 0.5 probability).

The RoBERTa White supremacy model classified the texts by John Earnest and Anders Breivik as white supremacy. The Random forest model classified the texts by John Earnest, Anders Breivik, Brenton Tarrant, Peyton Gendron, and Dylan Roof as white supremacy.

Both models classified the GP Animal Protection Manifesto and the Labours Animal Welfare Manifesto as animal rights and Greta Thunberg's speech as environmentalism and GP Animal Protection Manifesto and the Labours Animal Welfare Manifesto as environmental. The RoBERTa animal rights model also (incorrectly) classified Greta Thunberg's speech and Elliot Rodger's YouTube transcript as animal rights.

The RoBERTa incel model correctly classified Jake Davison's posts as incel but did not succeed in classifying Elliot Rodger's YouTube transcript as incel. The Random forest incel model correctly classified Elliot Rodger's YouTube transcript as incel but did not classify Jake Davison's posts as incel. The Random forest incel model incorrectly classified the texts written by John Earnest, Brenton Tarrant, Peyton Gendron, and Dylan Roof as incel (above 0.5 probability).

In summary, the results show that the performance of the models seems to differ when applied to new data.

VI. Discussion

Transfer learning using RoBERTa provided significantly better classification results than the model with Random forest and feature selection. However, the Random Forest model surprisingly worked better when applying the models to new unseen texts.

The performance of the different models differed depending on the different linguistic alignments. In the case study, the Random forest models worked much better on the right-wing alignments (counter-jihad, alt-right, and white supremacy) than the RoBERTa models. The RoBERTa models for Counter-jihad, Alt-right, and White supremacy did not perform well when it comes to classifying right-wing texts. The Random forest models could correctly classify five of the right-wing texts, while the RoBERTa models only managed to classify two texts (partially) correctly. However, it is important to note that the case study is small, and more data is needed to draw any conclusions from the results.

The RoBERTa incel model performed much better than the Random forest incel model. The Random forest incel model incorrectly classified four texts as incel, correctly identified one text as incel, and missed one text that should have been classified as incel. The RoBERTa incel model, on the other hand, missed one text and correctly classified one text as incel.

The small case study that we did only focused on longer texts. It would be interesting to use the models in a real scenario with shorter texts and get an understanding of the different model's performances on shorter texts. One of the challenges when using RoBERTa models is the limitation of the size of the texts that are classified. A RoBERTa model has a maximum token length of 512 tokens. The most common way to adhere to the limitations in text size is to only use the first 512 tokens, which is a sufficient option in many cases. Another option would be to split the text into multiple subtexts, classify each subtext and combine the results back together (for example, by choosing the class which was predicted for most of the subtexts). This latter option is more resource-consuming as all 512 token chunks in a long text must be classified.

VII. Conclusion and Directions for Future Work

We have examined the possibility of building classification models that can be used to determine linguistic alignment

TABLE V
THE RESULT OF THE CLASSIFICATION OF THE TEXTS USING THE ROBERTA MODELS (R) AND THE RANDOM FOREST MODELS (RF).

	CJ		**Alt-right**		**WS**		**Jihadist**		**Env.**		**AR**		**Incel**	
	R	RF	R	RF	R	RF	R	RF	R	RF	R	RF	R	RF
John Earnest	-	0.3	-	**0.8**	**0.98**	**0.9**	-	-	-	-	-	-	-	0.8
Anders Breivik	-	**1.0**	-	0.3	**0.98**	**0.9**	-	-	-	-	-	-	-	0.3
Anwar al-Awlaki	-	-	-	-	-	-	**0.8**	**0.99**	0.4	-	-	-	0.5	-
Brenton Tarrant	-	**0.9**	-	**0.5**	-	**0.8**	-	-	0.1	0.2	-	-	-	0.5
Peyton Gendron	-	**0.5**	-	**0.8**	-	**1.0**	-	-	-	0.2	-	-	0.4	0.8
Dylan Roof	-	0.4	-	**0.9**	**0.23**	**0.9**	-	-	-	-	-	-	0.1	0.9
Elliot Rodger	-	-	-	-	-	-	-	-	0.2	-	0.6	-	**0.3**	**0.99**
GP Animal Protection Manifesto	-	0.1	-	-	-	-	-	-	0.99	1.0	**0.99**	**0.9**	-	-
Greta Thunberg	-	-	-	-	-	-	-	0.1	**0.99**	**1.0**	0.91	0.1	-	-
Jake Davison	-	-	-	0.1	-	-	-	-	-	-	-	-	**0.99**	**0.1**
Labours Animal Welfare Manifesto	-	-	-	-	-	-	-	-	0.99	0.8	**0.99**	**0.9**	-	-

CJ = Counter jihad, WS = White supremacy, AR = Animal rights, Env. = Environmentalism, - = Result of the classification is 0

with a subculture, an online community, or an ideology. The results show that our linguistic alignment models seem to work well on the testing data. The models also performed relatively well on new unseen data. However, there was misclassification both when using the RoBERTa model and the Random forest model. The RoBERTa model incel model seems to work better than the Random forest model, but the Random forest model for the right-wing ideologies worked better than the RoBERTa models.

There are several directions for future research. One direction is to explore the possibility of using transfer learning and dividing longer texts into chunks and producing a combined classification result. Another direction is to explore transfer learning and models that can handle larger texts, such as Longformer [11]. Another direction is to improve the quality of the training data. We have mostly used entire discussion forums or subforums as training data, but the models may improve if the training data is chosen more carefully and annotated to ensure that the data is of high quality and representative of each linguistic alignment.

ACKNOWLEDGEMENTS

The research was supported by grants from Riksbankens Jubileumsfond to Nazar Akrami (P15-0603:1). The computations/data handling were enabled by resources provided by Chalmers e-Commons at Chalmers.

REFERENCES

[1] H. S. Alatawi, A. M. Alhothali, and K. M. Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374, 2021.

[2] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha. Detecting jihadist messages on twitter. In *European Intelligence and Security Informatics Conference*, pages 161–164, 2015.

[3] A. Berdicevskis and V. Erbro. You say tomato, i say the same: A large-scale study of linguistic accommodation in online communities. In *The 24rd Nordic Conference on Computational Linguistics*, 2023.

[4] K. Cohen and L. Kaati. Digital jihad. propaganda from the islamic state. In *Swedish Defence Research Agency FOI-R–4645–SE*, 2018.

[5] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words! linguistic style accommodation in social media. In *Proc of the 20th Int, Conf. on World Wide Web*, page 745–754, 2011.

[6] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: user lifecycle and linguistic change in online communities. In *WWW Conference*, pages 307–318, 2013.

[7] M. Dragojevic, J. Gasiorek, and H. Giles. *Communication Accommodation Theory*, pages 1–21. 12 2015.

[8] S. Jaki, T. D. Smedt, M. Gwóźdź, R. Panchal, A. Rossa, and G. D. Pauw. Online hatred of women in the incels.me forum. *Journal of Language Aggression and Conflict*, 2019.

[9] L. Kaati, K. Cohen, and B. Pelzer. *Heroes and scapegoats: right-wing extremism in digital environments*. European Commission and Directorate-General for Justice and Consumers., 2021.

[10] L. Kaati, E. Omer, N. Prucha, and A. Shrestha. Detecting multipliers of jihadism on twitter. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 954–960, 2015.

[11] R. Pappagari, P. Żelasko, J. Villalba, Y. Carmiel, and N. Dehak. Hierarchical transformers for long document classification, 2019.

[12] D. Parekh, A. Amarasingam, L. Dawson, and D. Ruths. Studying jihadists on social media: A critique of data collection methodologies. *Perspectives on Terrorism*, 12(3):5–23, 2018.

[13] B. Pelzer, L. Kaati, K. Cohen, and J. Fernquist. Toxic language in online incel communities. *SN Social Sciences*, 1, 2021.

[14] C. Sabater. Linguistic accommodation in online communication: The role of language and gender. *Revista Signos*, 50:265–286, 08 2017.

[15] N. Shuyo. Language detection library for java, 2010.

[16] A. Siegel, E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, and J. Tucker. Trumping hate on twitter? online hate speech in the 2016 u.s. election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1):71–104, 2021.

[17] S.Leader and P.Probst. The earth liberation front and environmental terrorism. *Terrorism and Political Violence*, 15(4):37–58, 2003.

[18] T. D. Smedt, G. D. Pauw, and P. V. Ostaeyen. Automatic detection of online jihadist hate speech. *CLiPS Tech. Report Series*, 7, 1-31, 2018.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Int. Conf. on Neural Inf. Processing Systems*, NIPS'17, page 6000–6010, 2017.

[20] S. Worchel, J. F. Morales, D. Páez, and J. Deschamps. *Social identity: International perspectives.* Sage Publications, Inc., 1998.

[21] M. Yoder, A. Diab, D. Brown, and K. Carley. A weakly supervised classifier and dataset of white supremacist language. In *Association for Computational Linguistics*, pages 172–185, 2023.