UNIVERSITY OF TWENTE.

Exploring the relationship between linguistic alignment and sentiment

In online discussions

Master's Thesis

Suzanna Wentzel
Interaction Technology

Supervisors

dr. Mariët Theune
dr. Doina Bucur
Sumit Srivastava, MSc.

Master's thesis

Suzanna Wentzel
Interaction Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente

Examination Committee

dr. Mariët Theune
Human Media Interaction
EEMCS-HMI
University of Twente

dr. Doina Bucur
Data Management & Biometrics
EEMCS-DMB
University of Twente

Sumit Srivastava, MSc.
Human Media Interaction
EEMCS-HMI
University of Twente

October 2023

Human-Media Interaction
Faculty of Electrical Engineering
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Acknowledgements

I would like to take this opportunity to thank some people to whom I owe my gratitude. First of all, I would like to express my thanks to my excellent supervisors: Mariët, Doina, and Sumit. I thank you for all of your wonderful support, sharing your expertise, and giving the pep talks when needed. I am very grateful for your extensive feedback (you have all read a lot and provided me with so many pointers), your infectious enthusiasm (about the topic, the details, the data, research in general, and cats), and the constructive meetings. I really enjoyed working with you.

Next, I would like to thank my study buddies. It was great to be working together, both online and in the Bastille, then in Technohal, and our final workspace Langerzijds. You have offered me great kindness, discussions, and company throughout the entire process. I'd like to especially thank Magdali for constantly reminding me to think whether steps were really necessary and keeping me on the right path and Joost for looking out for me when I got too stressed and teaching me so much about pandas.

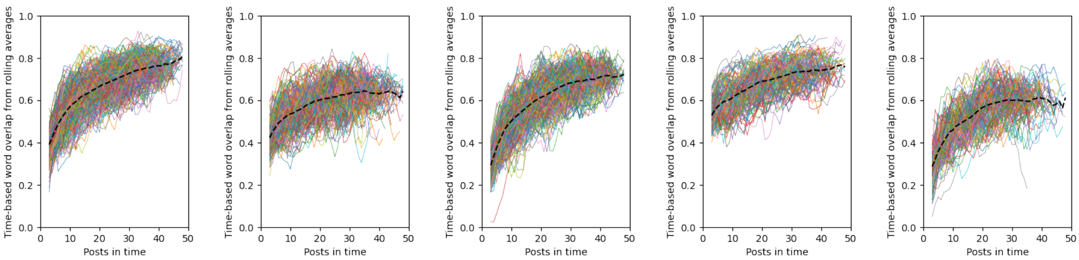
Finally, I want to thank my friends and family for supporting me, showing interest, believing in me, encouraging me, getting me out of the thesis bubble, checking in, and being there every other step of the way.

I have learned a lot from all of you, it would not have been possible without you.

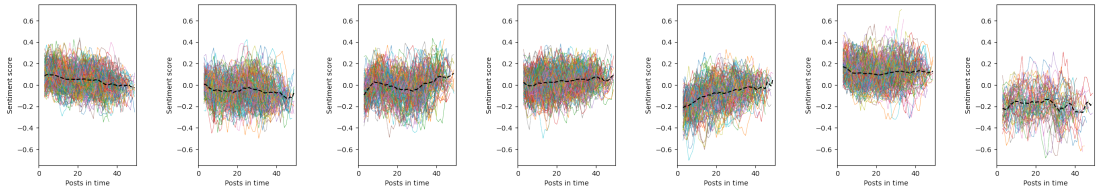
Summary

In conversations, people often converge on their linguistic behavior. This linguistic alignment has been researched in combination with many different factors, but a gap was found in studying the relationship between linguistic alignment and sentiment. This thesis aimed to find whether such a relationship exists, specifically in online multi-party political discussions on fora, to better understand human behavior. Specifically, the Internet Argument Corpus was investigated, which contains multi-party political discussions from 4forums with their annotated topic.

First, linguistic alignment on the lexical level was investigated. A literature study found that no suitable metric exists for analyzing multi-party alignment per post such that changes over time can be researched. Therefore, a new measure was proposed and applied: the time-based overlap. The dataset was analyzed to determine the distribution of time-based overlap of posts, the distribution of the average time-based overlap per discussion, and the time-based overlap over time. The alignment scores over time of discussions were clustered, resulting in groups with similar patterns of lexical alignment behaviors. Examples of different patterns are shown below (discussions with 30-50 posts):



Then, the sentiment per post was computed and analyzed. Similar to the alignment analysis, the sentiment distribution of posts, the distribution of average sentiment per discussion, and the sentiment scores over time were determined. The latter was clustered, resulting in groups with similar patterns of sentiment expression behaviors. Examples are shown on the next page (discussions with 30-50 posts).



Finally, the previous results were combined, and correlation and association coefficients were computed to find how sentiment and alignment relate. Contingency tables were computed to support the association findings between groups of sentiment and alignment patterns. Additionally, associations between alignment and sentiment, and discussion lengths and topic were investigated as the previous results hinted that these two factors might be involved.

Results showed no clear association between sentiment and alignment, but some association between alignment and topic was found, and even more association was found between sentiment and topic. For example, the topic “gun control” was found to be associated with opposing patterns of sentiments in discussions. Results also indicate that the length of discussions might be a factor. This implies that if sentiment and linguistic alignment do interplay, the correlation is complex and likely depends on other factors, such as this discussion length and topic.

Important limitations of this work are that the definition of the time-based overlap causes the scores to always increase, only one level of alignment is inspected, and topic words might influence the sentiment scores. Furthermore, only a limited amount of (longer) discussions were annotated with their topic, so we could not investigate all discussions in that regard.

Key recommendations are to use an improved and more recent dataset of discussions, to improve the time-based overlap measure by adding a penalty for the decay over time, and to investigate other factors that might influence the relation between sentiment and alignment.

To conclude, this research contributed to the field in several ways. An adapted metric for alignment was presented to measure it over time, which handles multi-party conversations and could function as a baseline for future work. Furthermore, we have found clear patterns of alignment and sentiment throughout discussions, meaning that such human behavior can be grouped and further investigated. Finally, we found that both alignment and sentiment appear to relate to the topic of the discussion.

Keywords

Computational linguistics, lexical alignment, sentiment analysis, multiparty online discussions

Contents

List of acronyms

ALIGN Analyzing Linguistic Interactions with Generalizable Techniques
BERT Bidirectional Encoder Representations from Transformers
CAT Communication Accommodation Theory
DTW Dynamic Time Warping
dynamic WHAM Dynamic Word-based Hierarchical Alignment Model
ER Expression Repetition
HAM Hierarchical Alignment Model
IAC Internet Argument Corpus
IAM Interactive Alignment Model
LLA Local linguistic alignment
LILLA Lexical Indiscriminate Local Linguistic Alignment
ML Machine Learning
NLP Natural Language Processing
POS Parts-Of-Speech
RepDecay Repetition decay
Scaled SCP Scaled Subtractive conditional probability
SCP Subtractive conditional probability
SILLA Syntactical Indiscriminate Local Linguistic Alignment
SWAM Simplified Word-Based Alignment Model
VADER Valence Aware Dictionary for sEntiment Reasoning
WHAM Word-Based Hierarchical Alignment Model

Chapter 1

Introduction

"Every language user, including young children and illiterate adults, can hold a conversation" - Pickering & Garrod

"If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart." - Nelson Mandela

The most natural and basic form of language use is dialogue, with the rise of social media allowing for dialogue to be held online. In conversations, people converge on behavior rapidly and unconsciously (Pickering & Garrod, 2006), it is something we all do. Recognizable situations might be when borrowing a word that a friend often uses, or using different words while talking to a boss or a colleague. These are examples of **linguistic alignment**: converging to conversational partners in communicative style and/or content (Xu & Reitter, 2015). It is a phenomenon that is more and more researched, in which its mechanisms are explored and different factors that could be related are found.

One factor that is not yet explored in combination with alignment is **sentiment**, expressing positive or negative opinions and emotions towards someone or something (Wilson, Wiebe, & Hoffmann, 2005). Expressing these sentiments is now widely possible on social media; a large portion of the content on social media contains opinions and sentiments (O'Connor, Balasubramanyan, Routledge, & Smith, 2010).

To gain a better understanding of human behavior, it would be interesting to see whether the two relate, as literature has already hinted at a possible relationship. Niederhoffer and Pennebaker hypothesize that speakers who are angry towards each other are highly likely to align their language use (Niederhoffer & Pennebaker, 2002). Bernhold and Giles state that alignment helps to communicate empathy (Bernhold & Giles, 2020).

Aside from adding to the knowledge of human behavior, knowing how alignment and sentiment relate might be useful for the Human Media Interaction field. If we for instance learn that we can stimulate positive sentiments by aligning linguistically in a certain way, the experience of users could be improved by adapting the language use of for instance conversational agents (e.g. social robots or chatbots), or including it in language models for text suggestions (e.g. autofill for chats or email).

This thesis explores how alignment and sentiment show in online discussions and investigates the interplay between the two. It also aims at overcoming additional challenges: including the context in measuring alignment and investigating multi-party conversations. We aimed to document this in a detailed way, such that others can reuse our methods. With this, we aim to shed light on human behavior so it can be used in other contexts as well.

The main research question is therefore the following:

> *How does linguistic alignment relate to the expressed sentiments of interlocutors in forum posts about political topics?*

With linguistic alignment, we mean the **lexical alignment**; alignment on a word level. Furthermore, for sentiments, we will look at the document level (overall sentiment in a post), in three classes (positive, negative, and neutral). Specifically, we will look at forum posts in online (multi-party) discussions about political topics on 4forums, an online forum for political debate and discussions. In the sub-questions, with "these discussions" we also refer to these discussions.

To be able to answer this question, we need to investigate linguistic alignment and sentiment, separately and in combinations. For both aspects (alignment and sentiment), we will first establish the big picture, finding the distribution of the aspects in the dataset. Then, we zoom in and investigate the finer patterns of the aspects changing over time.

Linguistic alignment

Q1.1 What is the distribution of the average lexical alignment in these discussions?
Q1.2 How does lexical alignment change over posts in these discussions?

Expressed sentiment

Q2.1 What is the distribution of sentiment expressed by interlocutors during these discussions?
Q2.2 How does the expressed sentiment change over posts in these discussions?

Interplay

Q3.1 How do the average alignment and the average sentiment of discussions relate?
Q3.2 How does the discussion length relate to the average alignment and the average sentiment in discussions?
Q3.3 How do the alignment and the sentiment trends over time of discussions relate?
Q3.4 How does the topic of a discussion relate to the sentiment and alignment trends over time of discussions?

Q3.2 and Q3.4 followed from the results of previous questions.

The remainder of this work is structured as follows. Chapter 2 explains the background of this thesis and relates it to literature. The dataset is investigated and described in Chapter 3 as more statistics were necessary for preprocessing the data for the next analyses. Then, lexical alignment is analyzed in Chapter 4 on average alignment per discussion, and over posts to find trends. Chapter 5 describes the sentiment analysis that investigates the average sentiment per discussion and over posts to find trends. Having the trends for lexical alignment and sentiment, their correlations are investigated in Chapter 6. Chapter 7 discusses the results on a higher level and suggests future work, and Chapter 8 compiles the answers to the research questions.

Chapter 2

Background

This chapter discusses the background of the study, covering communication accommodation, linguistic alignment, and sentiment analysis [1]. The final section discusses related work.

2.1 Communication accommodation

In conversations, conversational partners tend to adapt to each other in a broad range of behaviors. This adaptation is known as **communication accommodation**, which was described in the Communication Accommodation Theory (CAT) by Giles et al. (Giles, Coupland, & Coupland, 1991). Interlocutors attune their behaviors to meet the needs and desires of their conversational partner (Soliz & Giles, 2014), mostly non-consciously (Giles et al., 1991), but consciously is also possible (Bernhold & Giles, 2020). Soliz and Giles name different ways to attune, namely converging in communicative behavior, showing appropriate involvement, and using appropriate topics.

Most literature focuses on the converging (or diverging) behavior, which has been found in many different dimensions, such as posture (Condon & Ogston, 1967), pauses (Jaffe & Feldstein, 1970), gestures (Condon & Ogston, 1967) and linguistic style (Niederhoffer & Pennebaker, 2002). This converging behavior can thus both be verbal and nonverbal. It even exists in fictional dialogues, though it is stronger in actual interchanges (Danescu-Niculescu-Mizil & Lee, 2011).

Accommodation is ubiquitous, but individuals vary in levels of accommodation in ways that are socially informative (Doyle, Goldberg, Srivastava, & Frank, 2017). Interlocutors for instance converge their behavior in order to achieve social approval from in-group members of social groups, but diverge behavior in conversations with out-group members (Giles et al., 1991).

[1] This chapter is based on a literature study performed previously to this thesis by the same author.

2.2 Linguistic alignment

One important form of communication accommodation is **Linguistic Alignment** (Doyle, Yurovsky, & Frank, 2016). It is the primary form of accommodation in web-based communication (Wang, Reitter, & Yen, 2017). Different definitions of linguistic alignment are used, but in this thesis, we use linguistic alignment as the *act* of aligning/converging to conversational partners in style and content, as used by Doyle et al. and Xu and Reitter (2016; 2015), and the *resulting state* of having achieved alignment, as used by Pickering and Garrod (2004).

2.2.1 Levels of linguistic alignment

Linguistic alignment can occur in many different levels of language, such as the phonetic level, phonological level, lexical level, syntactical level, and semantical level (Pickering & Garrod, 2004). The last three levels will be explained further, disregarding the phonetic and phonological levels as they are related to the sounds of speech and not written text (Carrick, Rashid, & Taylor, 2016), and this research focuses on textual online conversations.

Lexical alignment is the alignment (repetition) of words (Pickering & Garrod, 2004) or word categories (Doyle et al., 2017). An example of lexical alignment of words, based on an example by Dubuisson Duplessis et al. (2021), is the repetition of the word "bunny" in the following:

A: Can you tell me something about the character, the white bunny?
B: The bunny is being chased by Alice.

Examples of word categories are articles, conjunctions, pronouns, negations, prepositions, quantifiers, etc (Doyle et al., 2016). An example of lexical alignment in word categories, in this case pronouns and prepositions, is given by Doyle and Frank:

A: I like **to** cook.
B: We love **to** eat.

Note that in this example, there could also be lexical (word) alignment in using the specific word "to" (Doyle et al., 2016).

Alignment at the **syntactical level** is the repetition of phrasal categories (Pickering & Garrod, 2004). It focuses on how people converse rather than on the content (Brinberg & Ram, 2021). An example is the following, where the same sentence structure is repeated:

A: I gave her the book
S(NP(PRP(I)), VP(VBD(gave), NP(PRP(her)), NP(DT(the), NN(book))))
B: And she gave him the summary
S(CC(And), NP(PRP(she)), VP(VBD(gave), NP(PRP(him)), NP(DT(the), NN(summary))))

Alignment at the **semantic level** is how the meaning of the utterance (in terms of what it tries to convey) aligns (Carrick et al., 2016). An example is given by Pilehvar, Jurgens, and Navigli (2013), where both utterances convey that an employee was fired by their manager:

A: A manager fired the worker
B: An employee was terminated from work by his boss

2.2.2 Properties of linguistic alignment

Many properties of alignment have been described in the literature. Properties of alignment as discussed in the Interactive Alignment Model (IAM) (described in Section 2.2.3) are that the process is primitive, ingrained in human behavior, occurs rapidly, and that it is mostly non-conscious (Pickering & Garrod, 2006) and automatic (Pickering & Garrod, 2004, 2006) (meaning that there is no conscious decision process followed for the act of aligning). Pickering and Garrod do note that it is possible to select expressions in a more intentional way (Pickering & Garrod, 2006) and that interlocutors could achieve alignment through explicit negotiation, but they normally do not (Pickering & Garrod, 2004).

Other properties that have been found are that linguistic alignment is partner-specific (Doyle & Frank, 2016). Doyle et al. note that both alignment and accommodation are usually incomplete; people become more similar but not the same (Doyle et al., 2016). Alignment was found to be present in different settings, such as in-person (Niederhoffer & Pennebaker, 2002) and web-based conversation (Danescu-Niculescu-Mizil, Gamon, & Dumais, 2011). Speakers with negative feelings towards each other are highly likely to coordinate linguistic style, while people who are not engaged are less likely to align (Niederhoffer & Pennebaker, 2002). Furthermore, alignment decreases over time (Wang, Reitter, & Yen, 2014).

2.2.3 Mechanism behind linguistic alignment

As Wang et al. write, the mechanism behind linguistic alignment is still debated (Wang et al., 2017). Two main theories have been presented and set out against each other in the context of linguistic alignment (see for instance (van der Pol, Gieske, & Fernández, 2016; Doyle & Frank, 2016)): the Interactive Alignment Model (IAM) and the Communication Accommodation Theory (CAT). The main theory that tries to explain linguistic alignment is the IAM. It describes that alignment is achieved by a parity between comprehension and production, where one interlocutor adapts their reference models based on the utterances of the other and the other interlocutor does the same, causing convergence over time, leading to more similar representations (Pickering & Garrod, 2006). This happens through three processes: using priming, repair mechanisms, and explicit other-modeling. Alignment at one level leads to more alignment at other levels (Pickering & Garrod, 2004, 2006), but the strength of the alignment may vary between levels (Duran, Paxton, & Fusaroli, 2019).

The second theory is the CAT, which focuses more on external factors influencing linguistic accommodation (for instance the wish to build rapport, a social goal).

2.2.4 Linguistic alignment measures

The main goal of linguistic alignment measures is to quantify the amount that one person's language use is influenced by another's (Doyle et al., 2016). Many measures have been presented and used in previous literature.

Local linguistic alignment (LLA)

Local linguistic alignment (LLA) originates from Fusaroli et al. (2012) and was adapted to specific LLA functions for lexical and syntactical alignment by Wang et al. (2014). LLA is a probabilistic (conditional) measure of alignment (by-word) (Doyle & Frank, 2016). There are two versions: discriminate (only counting words from a particular category) and indiscriminate (counting all words) (Doyle et al., 2016). LLA can measure two levels of alignment: on the lexical level with Lexical Indiscriminate Local Linguistic Alignment (LILLA), and on the syntactical level with Syntactical Indiscriminate Local Linguistic Alignment (SILLA).

Properties of LLA are the following. Discriminate LLA meets marker separability, meaning that alignment for different markers (words or word categories) can be evaluated separately (Doyle et al., 2016). Indiscriminate LLA does not meet marker separability (Doyle et al., 2016). It can be used to compute alignment in message pairs but does not incorporate message history.

Drawbacks of LLA are that it does not separate homophily from alignment, it is not consistent across different message lengths which means that it can be affected by reply length (Doyle et al., 2016) and it does not include order information, nor incorporates decaying over time (Doyle & Frank, 2016).

It was applied to a Twitter dataset (of 122693 message pairs between 2815 users) by Doyle et al. to compare different measures (Doyle et al., 2016). Carrick et al. used it on a smaller dataset of 103 dyadic conversations to see how well measures fit the IAM(Carrick et al., 2016). Xu and Reitter applied it to four different corpora: two online forums (Cancer survivor network and a massive open online course) and two published corpora (Switchboard and British National Corpus), also to compare measures (Xu & Reitter, 2015). The datasets of Doyle et al. and Xu and Reitter had potentially more than two interlocutors, the one of Carrick et al. had conversations with only two interlocutors.

Sequential pattern mining framework

Dubuisson Duplessis et al. propose a framework for computing various lexical alignment measures (Dubuisson Duplessis et al., 2021), which is available on Github[2]. It computes alignment from the recency adaptation perspective: it depends mainly on history. It is based on sequential pattern mining and focuses on two types of lexical patterns occurring during dialogue utterances: shared lexical patterns that are used by both interlocutors and self-repetition patterns that are used by one interlocutor.

[2]Framework: `https://github.com/GuillaumeDD/dialign`

The framework automatically builds a shared expression lexicon and self-repetition lexicons for both interlocutors. Shared expressions are established if they are produced by both interlocutors and occur at least once in a free form (when they are not dependent on another syntactic segment). Self-repetition expressions are established if the interlocutor used the expression at least twice and again if it occurs at least once in free form. An expression can be both in the shared expression lexicon and in a self-repetition lexicon. Subparts of expressions are discarded. With these lexicons, Duplessis created various measures for *strength*, *variety*, *complexity*, *stability*, and *alignment orientation*. One particular measure might be useful, the Expression Repetition (ER): proportion of tokens which interlocutors dedicate to the repetition of shared expressions.

Its main drawback is that the measures are designed for conversations between only two interlocutors.

Analyzing Linguistic Interactions with Generalizable Techniques (ALIGN)

ALIGN is a Python package that can measure different levels of linguistic alignment (lexical, syntactical, and semantical) in conversations (Duran et al., 2019). It extracts n-gram occurrences of two consecutive messages and converts them into vectors. Alignment is then measured by computing the cosine similarity between these vectors.

Its main drawbacks are that the tool only accepts input from conversations between two interlocutors and that they do not incorporate the conversation history for computing alignment.

Other measures

Another lexical measure is Subtractive conditional probability (SCP) (Danescu-Niculescu-Mizil et al., 2011), but it has the drawback that it does not take utterance length into account. Scaled Subtractive conditional probability (Scaled SCP) tries to overcome this drawback (van der Pol et al., 2016) but both are computationally complex and heavy. The Zelig Quotient measures the variance compared to the baseline use of lexical elements (sums of words in word categories) (Jones, Cotterill, Dewdney, Muir, & Joinson, 2014) and is a distributional measure of alignment. Its main drawback is that it computes similarity, not necessarily true alignment (Doyle & Frank, 2016). Lexical Similarity is a probabilistic (conditional) measure of alignment (Healey, Purver, & Howes, 2014). Its main drawback is that it is affected by message length and can be biased by the baseline frequency of the marker used (Doyle & Frank, 2016). Other measures for the lexical level are Lexical Repetitions (Ward & Litman, 2007) and Linguistic Style Matching (Niederhoffer & Pennebaker, 2002)

A measure that works for both the lexical and the syntactical level is Repetition decay (RepDecay) which investigates how much alignment decays over time (Reitter, Keller, & Moore, 2006). Its main drawback is that it cannot compute alignment on a turn-by-turn basis (Xu & Reitter, 2015). Another measure that works on both these levels is Spearman's Correlation Coefficient, which measures document similarity based on word frequency and co-occurrence (Kilgarriff, 2001). It has been found however that RepDecay and LLA perform better (Xu & Reitter, 2015).

A syntactical measure is Syntactic Similarity (Healey et al., 2014). Syntactic Similarity is another probabilistic measure of alignment where similarity is computed between each turn, in a similar way as for Lexical Similarity, but it works with syntactic structures instead of words.

Measures for the semantical level of alignment are deep learning semantic alignment as described by Xu (2021), and more simple word vectors that can be used to map words to a vector space and compare the semantical distance between the words (Carrick et al., 2016). Deep learning semantic alignment, more particularly Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019), might be more promising than such simple word vectors as it can capture the context and complex word characteristics.

Another set of promising measures is based on the Hierarchical Alignment Model (HAM), though these methods are complex. HAM is a probabilistic (conditional) measure of alignment focusing on function words (Doyle et al., 2016), proposed to overcome a lack of consistent measures across studies of linguistic alignment. It has been expanded upon in later works, such as the Word-Based Hierarchical Alignment Model (WHAM) (Doyle & Frank, 2016), Dynamic Word-based Hierarchical Alignment Model (dynamic WHAM) (Doyle et al., 2017), and the Simplified Word-Based Alignment Model (SWAM) (Shin & Doyle, 2018).

2.3 Sentiment analysis

Sentiment analysis is the task of identifying and extracting positive and negative opinions, emotions, and evaluations (Wilson et al., 2005), often towards some object (Jurafsky & Martin, 2023). Sentiment can be seen as a text classification problem (Bhagat & Bakariya, 2022), either with two classes (positive, negative) or three classes (positive, negative, and neutral) (Hung & Alias, 2023). Wilson gives the following examples of positive (+) and negative sentiments (-) expressed in sentences (Wilson et al., 2005):

- African observers generally approved (+) of his victory while Western governments denounced (-) it.
- A succession of officers filled the TV screen to say that they supported (+) the people and that the killings were "not tolerable (-)".
- "We don't hate (+) the sinner", he says, "but we hate (-) the sin".

There are various levels on which sentiment can be measured. Medhat describes the following (Medhat, Hassan, & Korashy, 2014):

- Document level (sentiment of a document, e.g. an entire post)
- Sentence level (sentiment of a sentence in a document, similar to document level but on a smaller scale)
- Aspect level (sentiment with respect to certain entities or properties of entities)

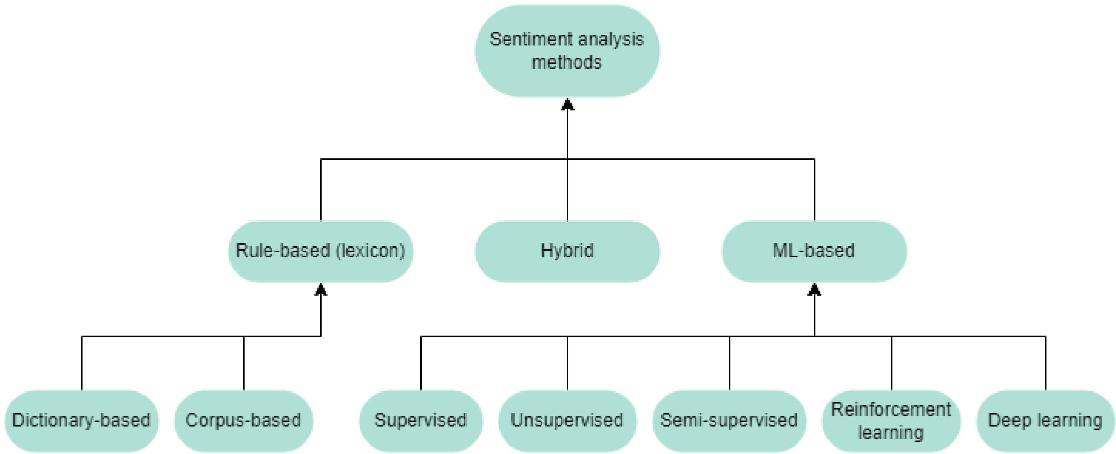


Figure 2.1: Sentiment analysis methods as classified in literature

Many methods exist for how to automatically extract sentiment from texts. Different papers maintain different groupings of methods (Jurafsky & Martin, 2023; Grover, 2022; Liu, 2010; Hung & Alias, 2023; Shobha Rani & Subramanain, 2020), which have been compiled in Figure 2.1. These methods can generally be categorized as rule-based or Machine Learning (ML)-based, but hybrids between the two also exist.

Rule-based methods use lexicons to extract sentiment from texts. Sentiment lexicons are lists of words that carry a strong sentiment, either positive or negative (Jurafsky & Martin, 2023). In the simplest method, as described by Jurafsky and Martin, these lexicons can be used to count the occurrences of positive and negative words (Jurafsky & Martin, 2023). Then, a text can be classified based on which has more occurrences. Often, a threshold is added, to only classify a document as positive or negative if a threshold is met. Another example of a rule-based method is Valence Aware Dictionary for sEntiment Reasoning (VADER), which combines rules that impact the intensity of the sentiment and a lexicon, and is attuned to microblog contexts (Hutto & Gilbert, 2014).

Challenges of using lexicon-based approaches are that the lexicons have difficulties detecting nuanced emotions or sarcasm and that they can depend on domain-specific characteristics (Grover, 2022). Also, words that are not in a lexicon might reduce performance (Wang, Guo, Yuan, & Li, 2022). Adding to that, in methods such as the rule-based method described above, from Jurafsky and Martin (2023), negation is not taken into account (though it is included in VADER).

Within the machine learning methods for sentiment analysis, there are five different subgroups of methods: supervised, unsupervised, and semi-supervised machine learning, reinforcement learning, and deep learning. The general steps in ML-based sentiment analysis are to collect the data, preprocess it (such as stemming, tokenization, stopword removal, spelling correction, etc.), extract features, train and then fit the model on the data (Bhagat & Bakariya, 2022). Features can for instance be n-grams, Parts-Of-Speech (POS) tags, or negations. Other features that can be useful for sentiment analysis are term frequency (Medhat et al., 2014), and emojis (Grover, 2022).

Supervised machine learning is useful when training data is available on which classifiers can be trained (Grover, 2022). Unsupervised methods do not need labeled data and are often simpler and faster (Grover, 2022). The goal of unsupervised methods is to find implicit patterns (Bhagat & Bakariya, 2022). Unsupervised methods can include a lexicon feature, where sentiment lexicons are used to count words that carry a sentiment (Grover, 2022). Semi-supervised machine learning focuses on both labeled and unlabeled documents. In reinforcement learning, the model learns by getting rewards for correct results and penalties for incorrect results, without supervision (Bhagat & Bakariya, 2022). Deep learning can be used for both document-level (Giachanou & Crestani, 2017) and aspect-based sentiment analysis (Shobha Rani & Subramanain, 2020). The advantage of using deep learning techniques is that these models learn the representations of the data by themselves, so feature selection is not necessary (Giachanou & Crestani, 2017).

Multiple machine learning methods can also be combined to overcome the drawbacks of individual classifiers (Bhagat & Bakariya, 2022), as such combinations generally achieve better results than using an individual one (Wang et al., 2022).

Machine learning approaches have several limits (Wang et al., 2022): the size of the training set has a significant impact on the classification performance, creating labeled training sets is a big task and ML-based approaches can be domain-dependent. Furthermore, they can be computationally expensive and take long to train or classify (Hutto & Gilbert, 2014).

Hybrid approaches address the weaknesses of both lexicon and machine learning based techniques (Wang et al., 2022). This increases performance but also requires a high computational cost (Giachanou & Crestani, 2017).

Ribeiro et al. compared many sentiment methods (lexicon, hybrid, and machine-learning based, including one deep learning method) on various corpora and found that for three-class sentiment analysis, VADER ranks the best (Ribeiro, Araújo, Gonçalves, André Gonçalves, & Benevenuto, 2016). It must be noted that it does not perform well on all datasets. SentiStrength performs best on two-class sentiment analysis. They recommend carrying out a preliminary evaluation, but if that is not possible, they recommend using one of the nine best-performing methods: Sentiment140 (Go, Bhayani, & Huang, 2009), Semantria (Lexalytics, 2015), OpinionLexicon (Hu & Liu, 2004), LIWC15 (Pennebaker, Boyd, Jordan, & Blackburn, 2015), SO-CAL (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), AFINN (Nielsen, 2011), VADER (Hutto & Gilbert, 2014) and Umigon (Levallois, 2013).

Pinto and Rocio describe several tools and APIs for doing sentiment analysis on the document level (2019), such as Amazon Comprehend (Services, 2023), Google Natural Language API[3], IBM Watson Natural Language Understanding API (Vergara, El-Khouly, Tantawi, Marla, & Sri, 2017) and the Azure Cognitive Service for Language[4]. However, the methods used for obtaining the sentiment are not transparent for any of these methods and most are paid or limited services.

[3] Google Natural Language API: `https://cloud.google.com/natural-language/`

[4] Azure Cognitive Service for Language: `https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/overview`

There are several challenges to be solved in sentiment analysis, such as dealing with sarcasm and ambiguities, incorporating context, figuring out the right features, and dealing with the difference between formal written texts and informal written texts (Hung & Alias, 2023). Social media texts do not follow standard language patterns and are considered informal, while most Natural Language Processing (NLP) tools are created with formal language in mind. Characteristics specific to social media are acronyms, redundant repetitions, no proper starting and ending of sentences, and site-related markup such as hashtags, labels, and tagging other users (Baldwin, Cook, Lui, MacKinlay, & Wang, 2013). Next to that, social media texts can contain noisy, unstructured data and misspellings (Shobha Rani & Subramanain, 2020).

2.4 Related work

2.4.1 Studies of alignment

Alignment has been investigated in discussion fora in combination with various factors such as power (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012), interpersonal relationships (Brinberg & Ram, 2021) and cultural fit within organizations (Doyle et al., 2017). Van der Pol et al. investigate linguistic style accommodation in the Internet Argument Corpus (IAC) and investigate its interplay with agreement and disagreement (van der Pol et al., 2016). They found that authors coordinate on style more noticeably if they disagree than if they agree. As they applied it to the same dataset that we have used, we could reuse their method. However, they compute the alignment only for message pairs, while we are interested in seeing alignment changing over time. We thus add to their research by using a different alignment metric and incorporating the entire message history.

Several studies investigated alignment with respect to time, but most of them are focused on investigating the decay effect, where they investigate the alignment against the distance between posts. One example is a study by Wang, Reitter, and Yen, which investigates multi-party conversations in online health communities (Wang et al., 2014), but investigates the alignment between message pairs and compares that with the distance between these messages. They do not investigate the alignment of all authors separately, but they do look, among other things, at alignment to the initial post-author. Another example is from Xu, who investigates semantic alignment between message pairs and compares that with the distance between the messages (Xu, 2021).

Duran performed an example study to show the performance of their new measure, ALIGN, where they looked at alignment over time (Duran et al., 2019). They compute alignment scores between contiguous turns. But as discussed before, their tool only works for conversations between two interlocutors and does not incorporate the message history.

In this thesis, we look at alignment over time, where we aimed to incorporate the context (history) in multiparty conversations. Existing measures were not sufficient for this situation, so we added to the field by inventing a new measure.

2.4.2 Studies of sentiment

There have been a plethora of studies that investigate sentiment. Here we describe two closely related studies that investigate sentiment over time in discussion data.

Kuzilek et al. investigated sentiment changes over time in a dataset from 5 discussion forums at a university, with discussions between students and teachers ranging from 2005 to 2009 (Kuzilek, Kravcik, & Sinha, 2020). Among other things, they investigated the sentiment trajectory to uncover the general sentiment trend and outliers. To calculate the sentiment, they used the SentimentWortschatz lexicon (Goldhahn, Eckart, & Quasthoff, 2012) to find the sentiments of individual message words and summed those for a message. Kuzilek et al. visualized the sentiment of messages in chronological order. They found that most forums tended to be slightly negative and that the trajectories had lowering sentiment values over time. They state that this implies that messages were more urgent towards the end. We add to this research by applying a similar method to a different dataset, and we use a more fitting sentiment analysis that is tailored to these microblogs. Furthermore, we inspected more than just sentiment as we tried to find clusters of patterns and used these patterns to find correlations with patterns in alignment.

Wen, Yang, and Rosé also investigate sentiment over time, but in their case to find the impact of sentiment on dropping out of courses over time (Wen, Yang, & Rosé, 2014). Similar to Kuzilek et al., they investigated sentiment expressed in posts of three course forums. They investigate the sentiment per day, meaning that they have to aggregate a lot of messages to find the public opinion on that day. They do this with collective sentiment analysis, where they find the topic sentiment ratio of positive versus negative words used in that day's post set with a sentiment lexicon. To get a more consistent trend, they apply a moving average over a window of the past k days. They found that the sentiment was much higher during the last course week. We are not investigating the public opinion, but the sentiment (changing) per discussion, so our analysis is slightly different. We have however reused the idea of the moving average.

2.4.3 Relating sentiment and alignment

There have been some hints to an interplay between sentiment and alignment in literature. Niederhoffer and Pennebaker hypothesize that speakers with negative feelings towards each other are highly likely to coordinate linguistic style, and people who are not engaged are less likely to align (Niederhoffer & Pennebaker, 2002). Berhold and Giles state that accommodation helps to fulfill affective goals, like decreasing social distance between interlocutors and communicating empathy or agreement (Bernhold & Giles, 2020), which would mean that alignment could lead to more positive engagement. However, to the best of our knowledge, no study has yet been performed to investigate this interplay between sentiment and alignment.

2.5 Conclusion

In this section, we have explained several concepts. **Communication accommodation** is when interlocutors adapt to each other to meet each other's needs, mostly by converging their behaviors. **Linguistic alignment** is a subset of communication accommodation and covers converging behavior to conversational partners in style and content. Linguistic alignment has many levels (from phonetic to semantic), and for each level, many methods have been proposed to measure the phenomenon. **Sentiment analysis** is the task of identifying and extracting positive, negative, and/or neutral opinions, emotions, and evaluations, often from text.

We have identified several gaps in existing works, such as measuring the alignment over time while incorporating context, measuring alignment in a multiparty context, and investigating the interplay between sentiment and alignment.

Chapter 3

Dataset analysis & preprocessing

This chapter describes the dataset, the methods of inspecting it further, the found dataset statistics and how those influenced the preprocessing of the data such that it could be used by the next analyses.

3.1 Dataset

The IAC was used because it meets the following requirements:

- It is a conversational dataset containing English online discussions.
- It (mostly) has multiple turns/posts per interlocutors, such that changes over time can be measured.
- The conversations are multiparty (though dyads would also be sufficient).

The Internet Argument Corpus (IAC) was also used in the related study by Van der Pol et al. (2016). It was created by Walker et al. (2012). A second version was published in 2016 (Abbott, Ecker, Anand, & Walker, 2016). The second version provides more data than the first, and it is now organized in an SQL schema (instead of JSON and CSV). The IACv2 contains dialogues from several fora: 4forums.com[1], CreateDebate.com[2] and ConvinceMe.net[3]:

- **4forums**: Website for political debate and discussions, covering topics relevant to the politics of the US. Posts are annotated with topic, agreement, and stance. It contains 11,079 threads with in total 414,453 posts.
- **CreateDebate**: Highly structured debate site with two parties. The dataset contains a subset of the discussions from CreateDebate, specifically focused on the topic of gun control. Posts are annotated with stance and topic. It contains 5413 debates with in total 65,368 posts.

[1] Now offline
[2] `https://www.createdebate.com/`
[3] Now offline

- **ConvinceMe**: Highly structured debate site. Posts are annotated for stance and topic. It contains 2958 posts.

The 4forums dataset contains the highest number of discussions and the highest number of posts, so the 4forums dataset was used. The 4forums dataset contains 11079 discussions with at least two posts (414.453 posts in total, by 3452 authors), of topics relevant to US politics. Posts are annotated with topic (2894 discussions), agreement (11980 posts), and stance (2248 authors by topic), among other things.

The text of posts in the dataset contains emoticons, capitalized letters, and URLs. Emojis are not present but are replaced by a plaintext description.

3.2 Dataset statistics

More statistics about the dataset were needed such that the data could be preprocessed correctly for the next analyses and to obtain information about the context of the dataset.

3.2.1 Time span of the posts

The time span of the posts was investigated to get insight into the context of the dataset. This was done by loading the dataset using SQL. The posts were ordered on creation date, and the first and the last were extracted to obtain the time span of the posts in the dataset. It was found that posts ranged from January 2003 until September 2012.

3.2.2 Number of discussions that are big enough

The number of discussions that are big enough was inspected to get insight into the size of the dataset. By "big enough", we mean a discussion with at least two authors who have at least four posts. With two authors having four posts, it is more likely that there is interaction between the authors resulting in alignment and that there are enough posts to see alignment changing over time.

After loading the dataset using SQL, the discussions that had at least two authors with at least four posts were extracted into a .csv and counted. 5424 discussions with at least two authors having at least four posts were found.

3.2.3 Number of posts per discussion

To get insight into the scope of the time series that could be applied and to check for outliers, the number of posts per discussion was inspected.

The extracted (big enough) discussions were ordered by discussion id and post id, and the number of posts per discussion was computed and inspected by computing the mean, the minimum, the maximum, and the 1^{st}, 5^{th}, 10^{th}, 50^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.5^{th} percentiles. Furthermore, the number of posts per discussion was plotted in histograms to get a visual overview. Potential outliers in tails or spikes were inspected by taking a random sample of these percentiles. A true discussion contains at least some statement that authors would provide arguments for and/or against.

Statistics on the number of posts per discussion (or length of discussions) are described in Table 3.1. The number of posts per discussion ranges between 8 (because we filtered on discussions with at least two authors with at least four posts) and 1291 posts. The median discussion length is 43 posts.

	Mean	**Min**	**1^{st}**	**5^{th}**	**10^{th}**	**50^{th}**	**90^{th}**	**95^{th}**	**99^{th}**	**99.5^{th}**	**Max**
posts	68	8	11	16	19	43	138	202	388	485	1291

Table 3.1: Statistics on discussion length (in number of posts)

The distribution of posts per discussion can be seen in Figure 3.1. The top figure shows the number of discussions per discussion length, the middle has a log scale on the y-axis, and the bottom figure zooms in on the first two bins of the top figure. One can see that the distribution of discussions on discussion length has a long tail, with very few discussions having many posts and more discussions having few posts. This visual observation is confirmed by the percentile statistics in Table 3.1.

When investigating a random sample of the discussions that make the tail of the last 0.5^{th} percentile, nothing stood out in particular. They were actually discussions, often about religion or abortion, which could be topics that people have very strong feelings about. Therefore, based on the discussion length, there was no reason to act on outliers.

3.2.4 Number of authors per discussion

Again to get insight into the scope of the time series that could be applied and to check for outliers, the number of authors per discussion was inspected.

The number of authors of the extracted discussions was investigated by computing the mean, minimum, maximum, and the 1^{st}, 5^{th}, 10^{th}, 50^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.5^{th} percentiles. Furthermore, the number of authors per discussion was plotted in histograms to get a visual overview. Potential outliers in tails or spikes were inspected by taking a random sample of these percentiles. The same criterium for defining a true discussion as before was used.

Table 3.2 shows the statistics on the number of authors per discussion. It ranges from 2 to 159 authors per discussion, with a median of 11 authors.

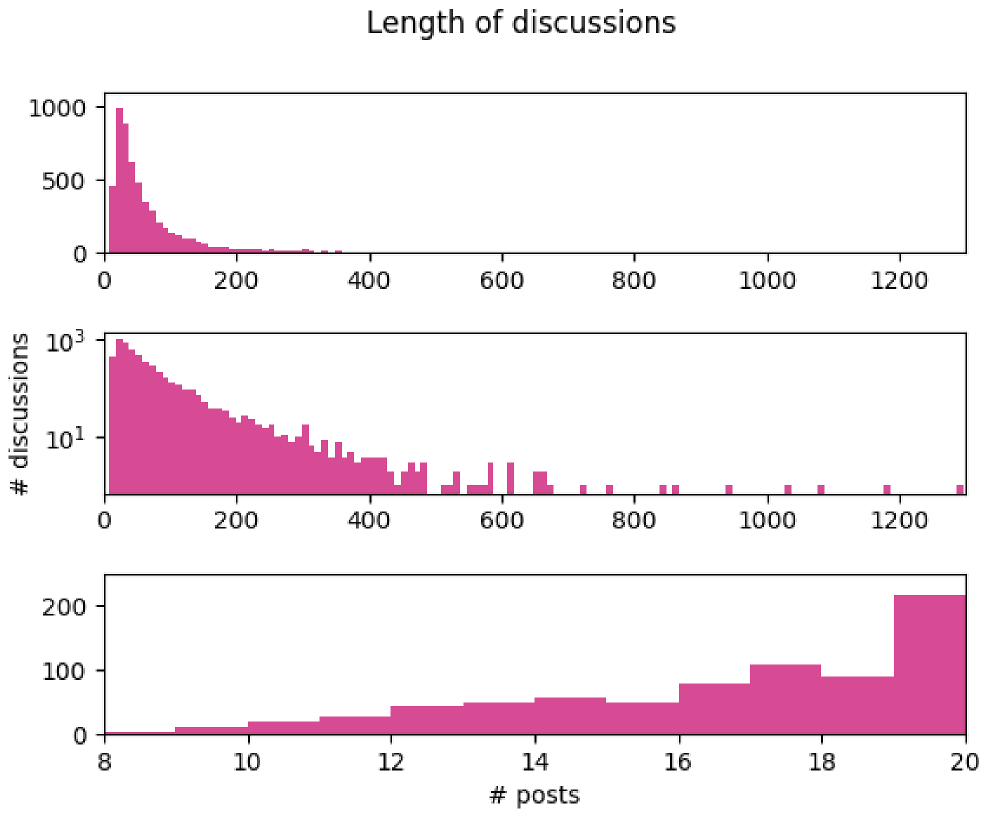


Figure 3.1: Histograms of discussion length

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	**90th**	**95th**	**99th**	**99.5th**	**Max**
Authors	12	2	3	5	6	11	21	25	39	46	159

Table 3.2: Statistics on the number of authors per discussion

The distribution of the number of authors per discussion is plotted in Figure 3.2. The distribution is bell-shaped, with a longer tail on the right. The distribution and the percentiles show that there are more discussions with fewer authors and fewer with more authors.

Inspecting the discussions with more than 60 authors showed that these were all discussions, with the common factor that they were quite long. Therefore, based on the number of authors per discussion, there was no reason to act on outliers.

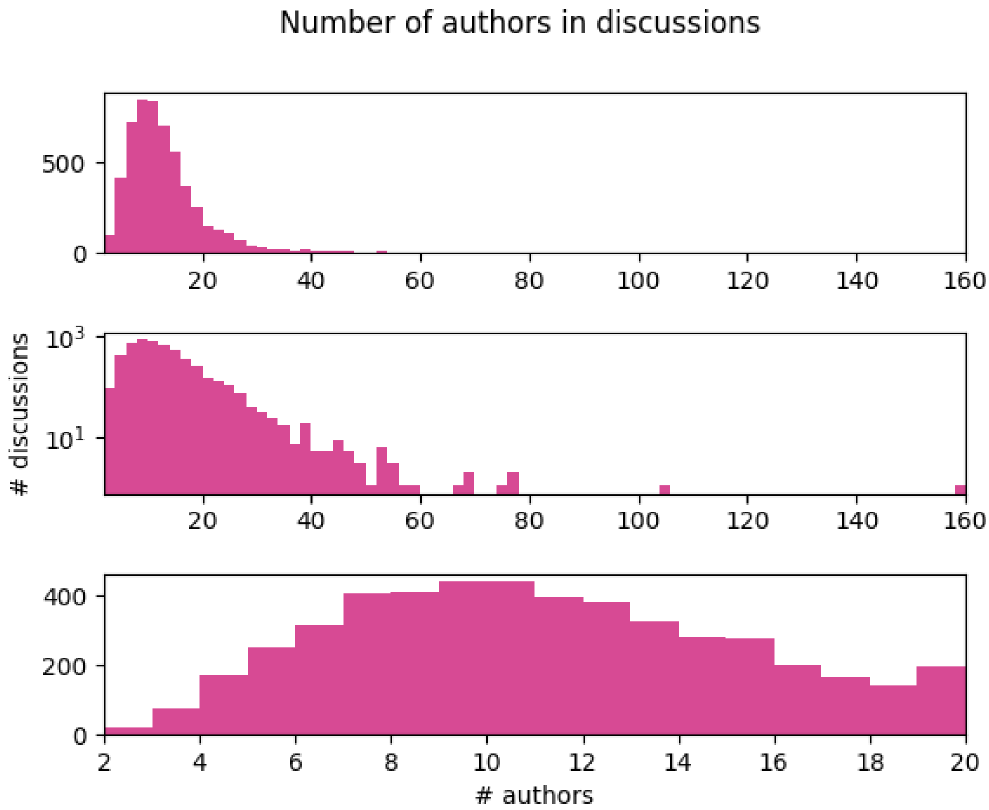


Figure 3.2: Distribution of authors per discussion

3.2.5 Author contribution

To get more insight into the scope of the time series that could be applied, the authors' contributions per discussion were inspected.

The mean, minimum, maximum, and the 1^{st}, 5^{th}, 10^{th}, 50^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.5^{th} percentile of posts per author in discussions were computed. In addition, heat plots were created to show the proportion of discussions with a certain number of authors and discussion length. Furthermore, for all discussions, the ten most contributing authors were found, and for each, the fraction of posts with respect to the discussion length was computed. Over all discussions, for each of these 10 most prolific authors, the mean contribution in the fraction of posts per discussion length was computed, to find how many authors should be included in the time series analysis, as well as the cumulative contribution

As shown in Table 3.3, the number of posts per author per discussion ranges between 1 and 285 posts, with a median of 2 posts. This means that there are many authors who contribute little to discussions.

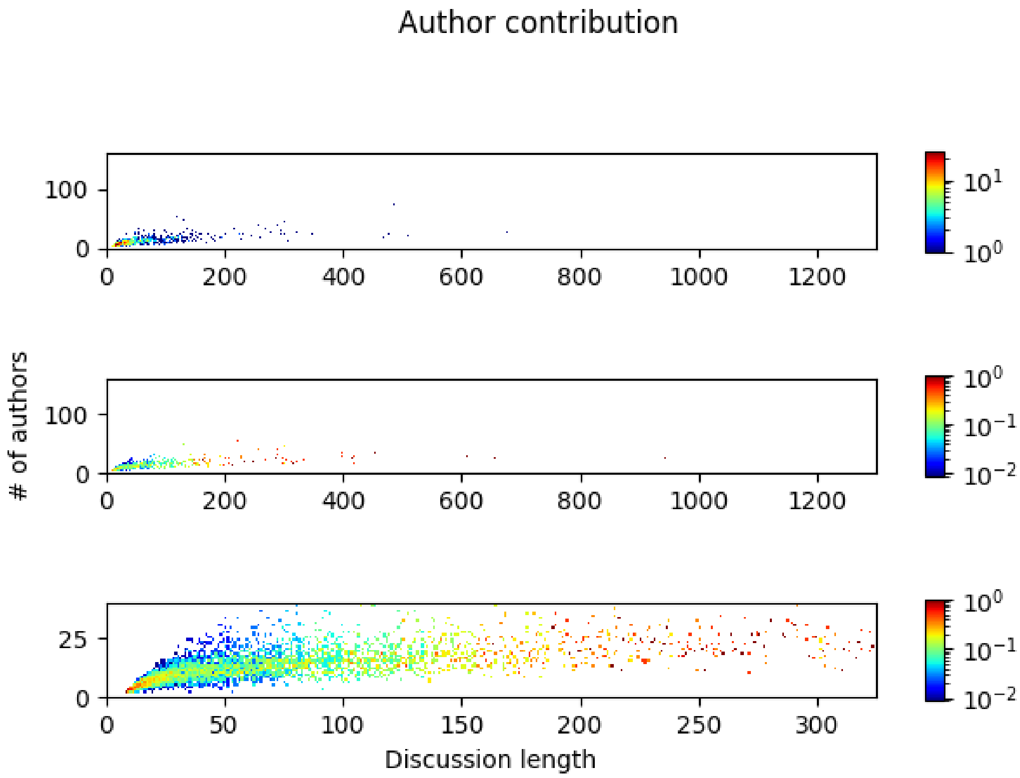


Figure 3.3: Number of posts per number of authors per discussion length

	Mean	Min	1st	5th	10th	50th	90th	95th	99th	99.5th	Max
Posts	5	1	1	1	1	2	13	19	42	56	285

Table 3.3: Statistics on posts per author per discussion

Figure 3.3 shows heat plots of the number of discussions per number of authors and discussion length (number of posts). The first graph in the figure shows in each cell the number of discussions per discussion length and number of authors. In the second and third graphs, each cell has been normalized by dividing by the total number of discussions with the corresponding discussion length, resulting in a proportion of discussions with that length per cell. The third graph is a zoomed-in version of the second. A clear trend is visible, where the amount of authors stabilizes as discussions get longer. So, in longer discussions, fewer authors are responsible for more posts compared to shorter discussions.

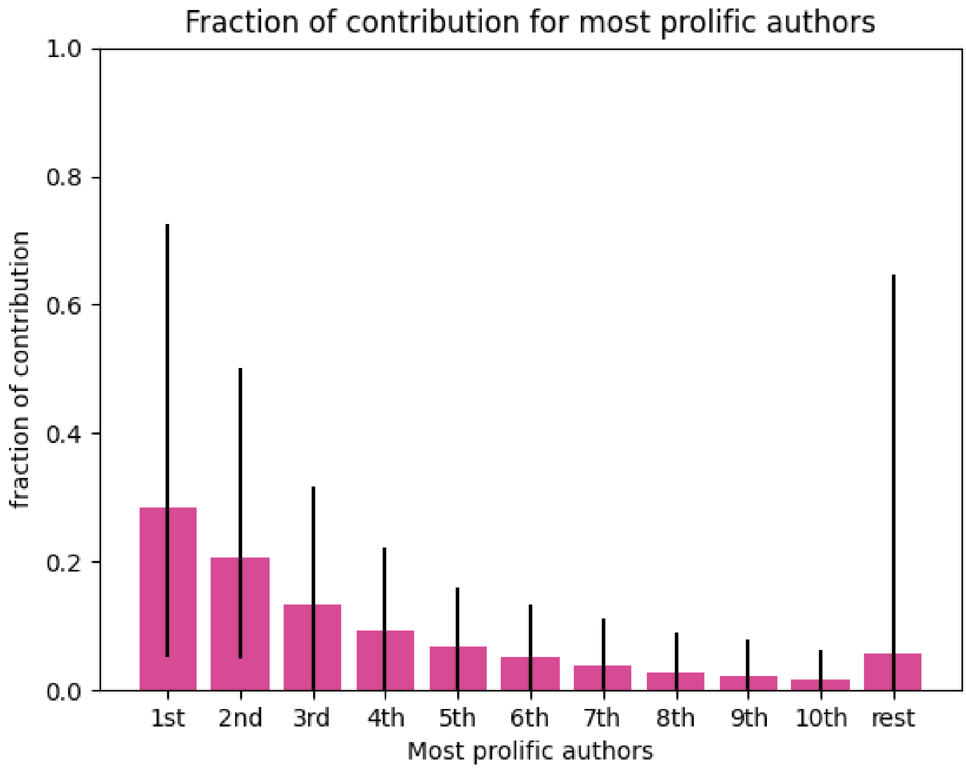


Figure 3.4: The mean contribution in fraction of entire discussions of the most prolific authors, with an indication of the minimum and maximum values

Prolific authors	*1st*	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*	*10th*	*Rest*
Contribution	0.285	0.206	0.134	0.094	0.068	0.051	0.038	0.029	0.022	0.017	0.057
Cumulative contribution	0.285	0.491	0.625	0.719	0.787	0.838	0.876	0.905	0.927	0.943	1.000

Table 3.4: Mean contributions per most prolific authors in discussions

A bar plot of the mean contribution (the number of posts contributed divided by the number of all posts in the discussion) of the most prolific authors per discussion is shown in Figure 3.4. The mean scores with the cumulative contribution are shown in Table 3.4. This table shows that on average, 8 authors account for more than 90% of the discussion.

3.2.6 Time-based overlap between posts in discussions

A manual inspection of the dataset showed that some of the discussions were not in fact true discussions. An initial inspection of the lexical alignment revealed these outliers. Therefore, the mean time-based overlap was found for all discussions (see Chapter 4 for more information on the measure). First, the discussions needed to be preprocessed (as described in Section 3.3 and 4.1, but without removing the outliers based on this analysis) because the time-based overlap can only be applied to preprocessed data. From the resulting overlap scores, the mean, minimum, maximum, and 1st, 5th, 10th, 50th, 90th, 95th, and 99.5th percentiles were computed, along with a histogram of the number of discussions per overlap score. From the 1st, the 50th, and the 99.5th percentile, random samples were extracted and investigated to see if they were discussions or not. In addition, potential outliers were identified from the histogram.

The statistics of the mean time-based overlap in the discussions are described in Table 3.5. The distribution of average overlap in the discussions is shown in Figure 3.5.

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	**90th**	**95th**	**99th**	**99.5th**	**Max**
time-based overlap	0.6021	0.2208	0.3466	0.4133	0.4587	0.6030	0.7476	0.7833	0.8373	0.8536	0.9838

Table 3.5: Statistics on the average time-based overlap in discussions

The distribution is bell-shaped, with a mean time-based overlap of 0.52. Two things stand out from the histogram: high spikes at the beginning, around an overlap below 0.3, and very high alignments at the tail, after 0.92.

Inspecting a random sample from 10 discussions out of the 1st percentile showed that one of them was not a discussion but a movie title ABC. The other conversations were in fact discussions, about war, abortion, evolution, and religion. Based on this sample, there was no reason to act on outliers.

Taking the discussions with an overlap below 0.3 showed that one of these five discussions was not a true discussion but a rhyme game. The other discussions were about wars, abortion, and a previous president of the US. Based on these results, there was no reason to act on outliers

Inspecting a random sample of 10 discussions around the median overlap of the discussions showed that all conversations are discussions.

Taking a random sample of 10 discussions with an overlap higher than the 99.5th percentile showed that one of the conversations was not a true discussion but a continuation story. The actual discussions were about religion, abortion, gay marriage, evolution, and propaganda. Based on this sample, there was no reason to act on outliers.

Investigating the discussions with an overlap higher than 0.92 showed that these were all not discussions but continuation story games. These outliers should be removed.

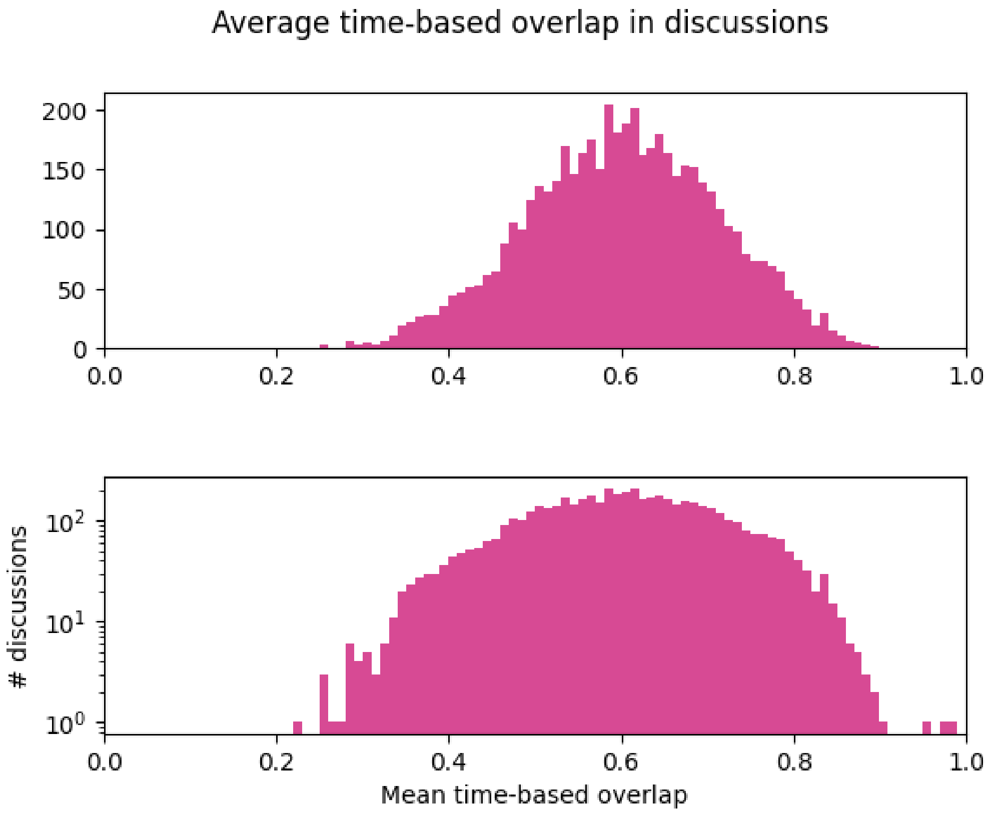


Figure 3.5: Histograms of average time-based overlap in discussions

3.3 Preprocessing

Based on the results of the dataset inspection, the discussion data was preprocessed such that the analyses for linguistic alignment and sentiment could be applied. Figure 3.6 shows the preprocessing steps and the resulting number of discussions. The steps are now further described.

First, we investigated discussions with at least two authors with at least 4 posts, as in such discussions it is more likely that there is an interaction between authors and there are enough posts to inspect changes over time (see Section 3.2.2). This reduced the dataset from 11079 discussions to 5424 discussions.

Then, the discussion threads were extracted. When the user interface of 4forums was inspected, we found that there are three modes of viewing a discussion: linear mode (viewing all posts ordered by date of posting), threaded mode (a tree structure with posts linked to their parent post), and a hybrid mode (a combination of the two), see Appendix A. Unfortunately, we could not find a snapshot of the hybrid view.

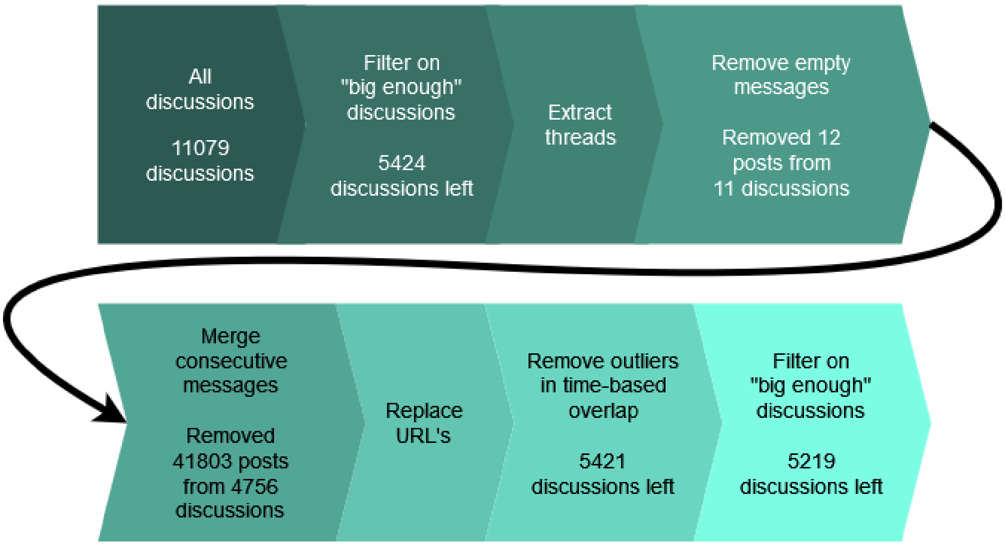


Figure 3.6: Dataset reduction & preprocessing steps

We cannot be sure which mode users used when joining the discussions, and it is very likely that both within and between users, these modes have been switched. However, to keep the scope of this study limited, we cannot investigate both. As authors could potentially be influenced by any of the posts that have been posted earlier in a thread, we looked at the linear threads. These were created by ordering the posts within discussions by creation date.

Eleven of the posts (all part of different discussions) of the remaining discussions did not contain any text. These messages were removed from the discussions (and their threads).

Adjacent messages of the same users in a thread were merged, as one could see them being in the same turn and we are interested in the linguistic alignment between interlocutors, not in self-alignment. 41803 posts of 4756 discussions were merged into previous messages of the same author.

For each of the messages, URLs were replaced by the tag *[URL]*, following the same strategy as Carrick, Rashid & Taylor (2016). Other than Carrick et al., screennames have not been replaced, as a quick manual inspection of the posts showed that different forms of the names were used, inconsistently.

Then, with the insights of the data statistics as presented in Section 3.2.6, 3 discussions with an average overlap higher than 0.9 were removed.

Because of removing empty messages and merging consecutive messages of the same authors, discussions could have been edited such that they do not contain at least two authors with at least four posts anymore. Therefore, another layer of filtering out discussions that do not make this criterion is applied. 202 of such discussions were removed. The resulting 5219 discussions are used throughout the next chapters.

The notebook used for preprocessing the data can be found at the Github repo[4], see "step 1: preprocessing" in the README.

[4]https://github.com/SuzannaWentzel/Sentiment-Alignment-Interplay

Chapter 4

Lexical alignment

This chapter discusses the first step in answering the main research question and investigates the lexical alignment in the discussions. It answers the following subquestions:

Q1.1 What is the distribution of average lexical alignment in the discussions?
Q1.2 How does lexical alignment change over posts in the discussions?

4.1 Preprocessing for lexical alignment

After the general preprocessing as described in Section 3.3, additional preprocessing was applied for the lexical alignment analysis. The steps are shown in Figure 4.1 and are further explained below.

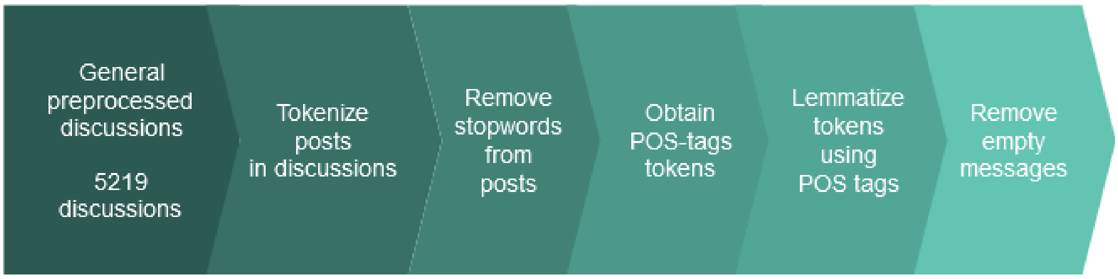


Figure 4.1: Additional preprocessing steps for analyzing lexical alignment

Posts were tokenized with `word_tokenize` from the `NLTK.tokenize` python package to obtain a list of words in the message. Each token was converted to lowercase.

Stopwords, words that carry little information because they are so commonly used (and are therefore not useful indicators of alignment), were removed, as they would make the scores appear higher than they actually are. The stopword list that was used came from the `NLTK.corpus` package.

Furthermore, we applied lemmatization, a process to obtain the lemmas (often the stem) of words, such that, for example, the following situations were also recognized as lexical word alignment:

Situation 1: word inflections
Message: I have a bunny.
Response: Ah, I love bunnies!

Situation 2: (some) word derivations
Message: I was working today.
Response: Did you have a lot of work to do?

Inflections are always captured with lemmatization because they should have the same lemma (though a lemmatizer can make mistakes). Derivations generally have different lemmas and are not seen as alignment, except for situations similar to situation 2, where the lemma of a word with one POS is the same as the lemma of another word with a different POS. In situation 2, "working" in the message is a verb and gets lemmatized to "work", whereas "work" in the response is a noun and also gets lemmatized to "work".

The lemmatization was done with `WordNetLemmatizer` from the `NLTK.stem` Python package. To be able to run the lemmatizer, words also had to be tagged with their POS. This was applied with the `pos_tag` function of NLTK. This POS-tagger was trained on the Penn Treebank corpus, and returns POS-tags from the Penn Treebank tagset. The `lemmatize` function expects tags from the Wordnet tagset, so these Treebank POS-tags that were obtained from the POS-tagger had to be transformed to Wordnet tags. With the Wordnet tags, lemmatization was applied.

After these preprocessing steps, messages could have become empty, if posts consisted entirely of stopwords. Therefore, we added another layer of preprocessing by removing empty preprocessed posts.

4.2 Alignment metric

There are many methods for measuring linguistic alignment. Here, we focus on the lexical alignment, and more specifically the word repetition (so not the word category repetition), which was explained in Section 2.2.1. Other levels were not investigated because of time constraints.

In the literature review performed prior to this study, many methods have been compared and it was found that LILLA (see Section 2.2.4) would be the most suitable method for evaluating lexical alignment. However, this metric is less useful for computing alignment over time, because it does not include any history or context of the words that have been used previously.

Therefore, we use a metric inspired by the Expression Repetition (ER) that was described by Duplessis et. al (Dubuisson Duplessis, Clavel, & Landragin, 2017). They describe ER as the ratio of produced tokens belonging to a repetition of an established expression:

$$\text{ER} = \frac{\#\text{Tokens in an established expression}}{\#\text{Tokens}}$$

An expression is established if it has been produced by both interlocutors and at least once in free form.

Our dataset includes (often) more than two interlocutors per conversation, so we adapted their metric. We keep track of a vocabulary of tokens that have been used before in that discussion per author. For each new message m by author A, we obtain the number of tokens in that message which can be found in the vocabularies of the other authors $V_{\neg A}$ used previously in the discussion and divide that by the total number of tokens in the new message:

$$\text{Time-Based Overlap}(m_A, V_{\neg A}) = \frac{\#\text{Tokens in } m \text{ that are in } V_{\neg A}}{|m|} \tag{4.1}$$

This gives a score between zero and one, where 0 indicates no overlap with the previously used tokens and 1 indicates a full overlap with the previously used tokens.

This metric is illustrated with the following example discussion (generated by ChatGPT):

Example 4.2.1: Discussion about cats
Message 1 (person A): Cats are the best!
Message 2 (person B): No way, dogs are superior!
Message 3 (person C): Actually, I think both cats and dogs have their merits.
Message 4 (person B): But dogs are loyal and affectionate
Message 5 (person A): Agree to disagree, cats will always be my favorite!

This discussion is preprocessed to the following sequences:

Example 4.2.2: Preprocessed discussion about cats
Message 1 (person A): (cat best !)
Message 2 (person B): (no way , dog superior !)
Message 3 (person C): (actually , i think cat dog merit .)
Message 4 (person B): (but dog loyal affectionate)
Message 5 (person A): (agree disagree , cat always favorite !)

To calculate the time-based overlap, we need the previous vocabulary for each message, excluding the author's own tokens:

Example 4.2.3: Vocabularies for computing time-based overlap

Message 1 (person A): (cat best !)

> ¬A: -
> ¬B: cat best !
> ¬C: cat best !

Message 2 (person B): (no way , dog superior !)

> ¬A: no way , dog superior !
> ¬B: cat best !
> ¬C: cat best ! no way , dog superior !

Message 3 (person C): (actually , i think cat dog merit .)

> ¬A: no way , dog superior ! actually , i think cat dog merit .
> ¬B: cat best ! actually , i think cat dog merit .
> ¬C: cat best ! no way , dog superior !

Message 4 (person B): (but dog loyal affectionate)

> ¬A: no way , dog superior ! actually , i think cat dog merit . but dog loyal affectionate
> ¬B: cat best ! actually , i think cat dog merit .
> ¬C: cat best ! no way , dog superior ! but dog loyal affectionate

Message 5 (person A): (agree disagree , cat always favorite !)

> ¬A: no way , dog superior ! actually , i think cat dog merit . but dog loyal affectionate
> ¬B: cat best ! actually , i think cat dog merit . agree disagree , cat always favorite !
> ¬C: cat best ! no way , dog superior ! but dog loyal affectionate agree disagree , cat always favorite !

Computing the time-based overlap for message 1 is not possible as that is the first message of the conversation. For computing the overlap for message 2, we find that the overlap in tokens between message 2 and ¬B is only "!". The message has length 6, so the time-based overlap at message 2 is $\frac{1}{6}$. For message 3 (length 8), we find that the overlap in tokens between message 3 and ¬C is ", cat dog", so the time-based overlap at message 3 is $\frac{3}{6} = \frac{1}{2}$. Following the same methods, message 4 has a time-based overlap of $\frac{1}{4}$, and message 5 of $\frac{2}{7}$.

The implementation of the time-based overlap can be found in the Github repo[1], see the README for a reference.

[1] https://github.com/SuzannaWentzel/Sentiment-Alignment-Interplay

4.3 Methods

A notebook with the code for the methods that follow can be found at the Github repo[2], see step 2: alignment analysis in the README.

4.3.1 Distribution of lexical word alignment in discussions (Q1.1)

To find the distribution of lexical alignment in the discussions, the time-based overlap was applied to all preprocessed messages in the discussions. The distribution of the alignment of all messages was plotted in a histogram, and the mean, min, max, and 1$^{\text{st}}$, 5$^{\text{th}}$, 10$^{\text{th}}$, 50$^{\text{th}}$, 90$^{\text{th}}$, 95$^{\text{th}}$, and 99$^{\text{th}}$ percentiles were computed.

Furthermore, the average was taken per discussion, a histogram of the distribution was plotted and the same statistics as previous were computed. This information about the time-based overlap distribution gives a first insight into how alignment manifests in our dataset.

4.3.2 Lexical word alignment changing over posts (Q1.2)

Using the computed time-based overlap, we first plotted time-based overlap over posts for all discussions. A rolling average with a window of 5 was applied to smoothen the time series and to be able to find the general trends rather than the exact trends. We disregard the first and the last time points where the window of 5 cannot be applied. The rolling average of the time-based overlap was plotted.

Then, time series clustering was applied to find trends in our sequential data with `TimeSeriesKMeans` from the `tslearn.clustering` package. K-means clustering tries to minimize the total distance between all objects in a cluster from their cluster center (Macqueen, 1967). A cluster is represented by the mean of its objects. It is one of the most used partitioning clustering algorithms, dividing unlabelled data into k groups (Aghabozorgi, Seyed Shirkhorshidi, & Ying Wah, 2015).

We use the Dynamic Time Warping (DTW) metric for cluster assignment. DTW searches for the optimal alignment between two time series, warping nonlinearly to match the two as best as possible (Müller, 2007). This warps entire sequences, not subparts.

Even though `TimeSeriesKMeans` with DTW can handle time series of different time lengths, an initial clustering showed that it mostly clustered discussions of the same length together. Therefore we split the discussions up into bins of discussion lengths. The number of posts per discussion was computed, and discussions were split into 5 bins with approximately the same number of discussions per bin based on discussion length. All discussions of one length have been put in the same bin, so bins do not necessarily have an equal number of discussions. The last bin was furthermore split up into three more, to ensure that bins do not range too much in discussion length (at most with a difference of 50 posts). Bins that contained less than 50 discussions were discarded to ensure that there were enough discussions to obtain clusters.

[2] `https://github.com/SuzannaWentzel/Sentiment-Alignment-Interplay`

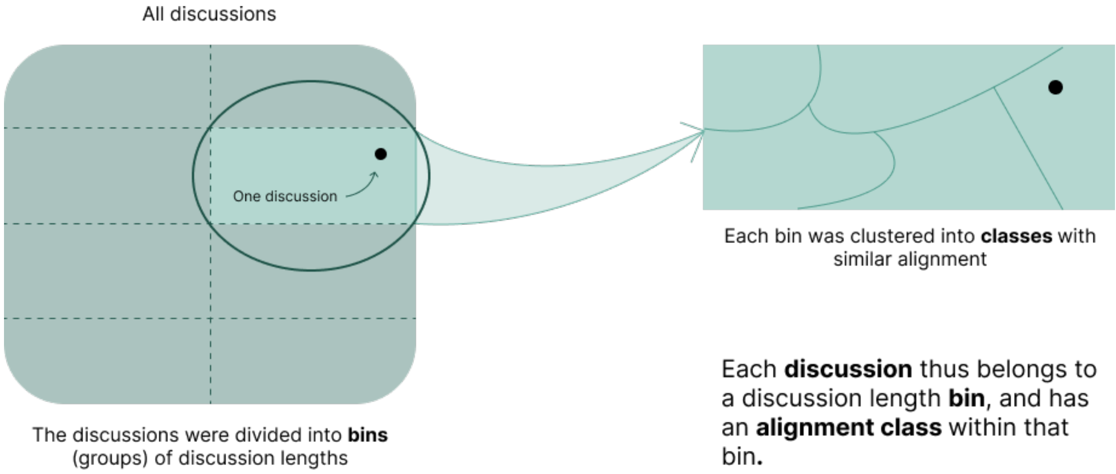


Figure 4.2: What is meant by bins and alignment classes

The number of clusters, k, was fine-tuned for each bin by running the `TimeSeriesKMeans` with k in the range $[1, 10]$ and fitting it on the time-based overlap for the discussions in that bin. One of the attributes that the model returns is inertia, the sum of the distances of samples to their closest cluster center. A lower inertia means that discussions in one class are more similar to each other than if there is a higher inertia.

This inertia was plotted against the number of clusters, and the elbow method was used to extract the optimal k per discussion length bin. The classes for all k for each bin were also plotted in line charts to support making the choice of the optimal k.

K-means clustering starts off with a random seed, which means that running the clustering several times with the same k returns different results. Therefore, we run the clustering for each bin with their optimal k five times and store the model with the lowest (best) inertia to obtain the most optimal clustering per discussion length bin.

For each discussion length bin, the optimal clustering was plotted per class, showing the different trends of time-based overlap over time (posts) per discussion length. Figure 4.2 can be used as a reminder of what is meant by *bin* and *alignment class*.

4.4 Results

4.4.1 Distribution of time-based overlap in discussions

The statistics of all time-based overlap of posts are shown in Table 4.1. The accompanying histogram is shown in Figure 4.3. These graphs show the number of posts per time-based overlap, where the second graph is a log-scaled version of the first, such that the lower number of posts per overlap does not disappear.

The distribution of time-based overlap of posts (Figure 4.3) is skewed to the left, with an average of 0.68 and median of 0.71. The highest number of posts has an alignment of around 0.8.

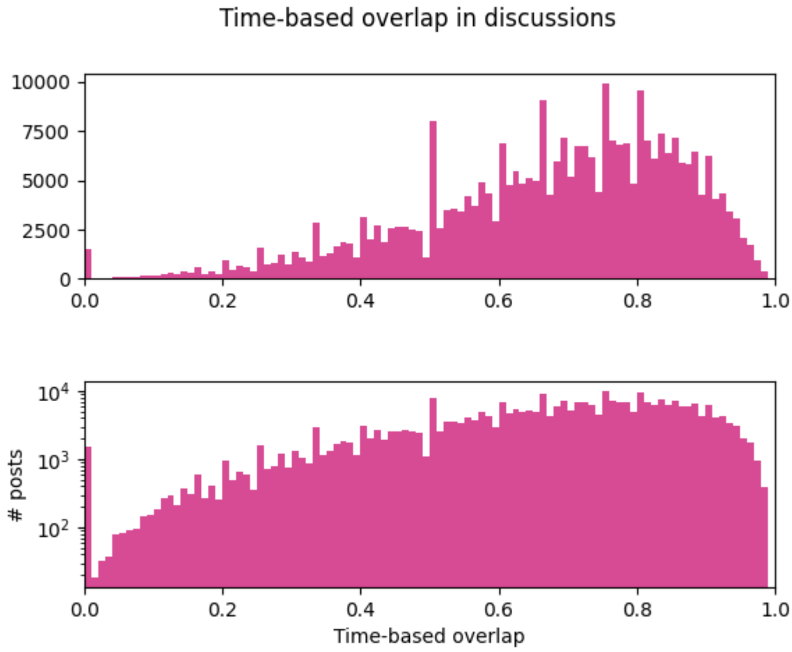


Figure 4.3: Distribution of time-based overlap of posts

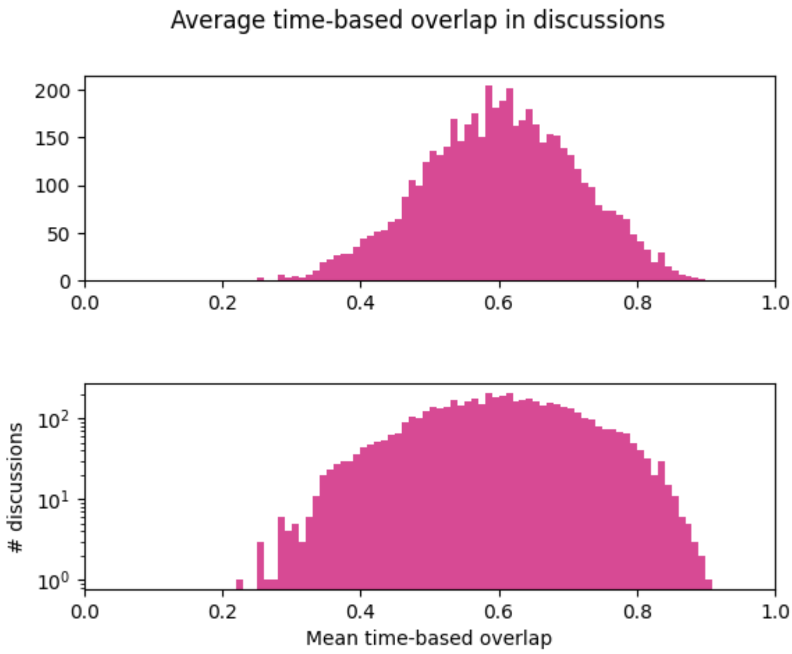


Figure 4.4: Distribution of average time-based overlap in discussions

	Mean	Min	1^{st}	5^{th}	10^{th}	50^{th}	90^{th}	95^{th}	99^{th}	Max
time-based overlap	0.6823	0.0	0.1379	0.3182	0.4107	0.7143	0.9054	0.9444	1.0000	1.0000

Table 4.1: Statistics of the time-based overlap of all posts

The statistics of the average time-based overlap in discussions are shown in Table 4.2. The accompanying histogram is shown in Figure 4.4. The graphs in the figure show the number of discussions per average time-based overlap. The second graph is a log-scaled version of the first, such that the lower number of discussions per time-based overlap does not disappear.

	Mean	Min	1^{st}	5^{th}	10^{th}	50^{th}	90^{th}	95^{th}	99^{th}	Max
time-based overlap	0.6019	0.2208	0.3466	0.4132	0.4587	0.6029	0.7470	0.7826	0.8356	0.9037

Table 4.2: Statistics on the average time-based overlap of all discussions

The average distribution (Figure 4.4) is obviously different from the distribution of time-based overlap of all posts. The distributions of posts are spread between 0.0 and 1.0, whereas the average distribution of discussions ranges between 0.2208 and 0.8356. Furthermore, the distribution of average alignment is not skewed but bell-shaped with an average of 0.6.

The average distribution also differs from the one seen in Figure 3.5, in that the outliers on the right have disappeared because they were removed.

An excerpt of a discussion (about abortion) with a characteristic average time-based overlap of 0.6 is the following, where the overlap is underlined:

Example 4.4.1: Discussion with characteristic alignment

...

Message 35 (author 2039): Then please explain the different <u>abortion</u> <u>laws</u> across the <u>EU?</u> (Time-based overlap: **0.44**)

Message 36 (author 1146): <u>Then</u> <u>please</u> <u>explain</u> why there is a <u>European</u> <u>Court</u> if it has no jurisdiction<u>.</u> I <u>don't</u> think we have any more to <u>say</u> to each other on this subject until <u>one</u> of us <u>supports</u> our position with web <u>links.</u> Otherwise<u>,</u> it is just guesswork<u>.</u> <u>I</u> prefer concrete <u>facts.</u> (Time-based overlap: **0.64**)

Message 37 (author 2039): <u>"It</u> is not the function of the <u>court</u> to tell individual <u>countries</u> what kind of detailed <u>laws</u> they should have on matters as controversial and sensitive as <u>abortion"</u> Barbara Hewson, barrister in human <u>rights</u> and <u>abortion</u> <u>law</u> BBC (Time-based overlap: **0.5**)

...

This example illustrates alignment in message 36, where the author starts with "Then please explain", repeating the words of the previous author. Author 1146 did however not align on using "EU" as author 2039 did, but typed it out as "European". It was still counted as alignment, as a previous author had used "European" before.

4.4.2 Finetuning clustering

The plots with alignment changing over posts for all discussions and the rolling average are shown in Appendix B. The plots finding the optimal cluster sizes per bin can be found in Appendix C.1. They plot the inertia per number of clusters. The optimal number of classes per discussion length bin derived from these figures are shown in Table 4.3.

Bin	Discussion length	# Discussions	Optimal k
1	7-22 posts	1089	4
2	23-33 posts	1049	6
3	34-50 posts	1058	5
4	51-86 posts	1052	5
5	87-136 posts	556	4
6	137-186 posts	174	5
7	187-236 posts	106	5
8	237-286 posts	52	4

Table 4.3: Optimal k per discussion length bin

Some of the elbows show some interesting behavior and do not follow the regular elbow shape as expected, for instance, the dip in bin 6 (Figure C.6), and the zigzagging in bin 8 (Figure C.8). Furthermore, the clustering for bin 7 (Figure C.13) showed one discussion having an alignment that stood out from the rest of the discussions.

Looking into the plotted classes of bin 6 (see Appendix C.2), for $k = 4, 5, 6, 7$, we see that with $k = 4$ and $k = 5$, there are distinct classes with high cardinalities. However, for $k = 6$ and $k = 7$, there are respectively one and two classes with lower cardinalities, and there appears to be more overlap in the classes (first and last graph of $k = 6$, second and last for $k = 7$). Having $k = 5$ resulted in a lower inertia than $k = 4$, so we chose $k = 5$ in bin 6.

Looking at the time-based overlap plots in bin 7 (see Appendix C.3), we see one discussion standing out from the others with a very low alignment, which also does not follow the mean trend. Investigating that specific discussion showed that it is not a true discussion, which has not been found in the preprocessing phase. It is actually a movie-title ABC.

Looking at the plotted classes for bin 8 (see Appendix C.4), for $k = 4$, $k = 5$, $k = 6$, $k = 7$, we see that for $k = 4$, it found distinct classes, though one with a lower cardinality. For $k = 5$, two classes have lower cardinality (first and fourth graph) and the first and the fifth appear to have some overlap. For $k = 6$ and $k = 7$, there are multiple classes with lower cardinalities and both have one class with only one discussion. Therefore, we chose $k = 4$ for bin 8.

4.4.3 Time series clustering on time-based overlap

The best performing time series clustering models for each bin are shown in Figures 4.5 - 4.17. The figures show the discussions per class that the clustering has found over time (posts). The figures differ in the number of classes, see Section 4.4.2.

For bin 1 (Figure 4.5), the four classes are different in shape and/or height of alignment. The third class is the steepest and ranges most in the time-based overlap. The fourth class has the flattest alignment. The first and the second have a similar shape where alignment increases, but the second starts off with a higher alignment than the first, with the first having the lowest alignment of all classes, and the second the highest of all classes.

Bin 2 (Figure 4.6) does not contain a class with as flat of an alignment as could be found in bin 1. However, classes 2, 4, and 5 have an alignment curve that is flatter than the rest of the classes in bin 2. Class 1 has the steepest increase of alignment, which is higher than that of the steepest class in bin 1. The alignment at the start of class 6 of bin 2 resembles the alignment of the third class of bin 1.

For bin 3 (Figure 4.8), the alignment of the third class seems to resemble class 6 of bin 2 and class 3 of bin 1. The alignment of the fourth class seems to resemble the fifth class of bin 2, and somewhat the second class of bin 1. The second class and the fifth class show an interesting trend of going slightly downwards at the end, which means a decrease in alignment.

This decrease is not present in the alignment trends for bin 4 (Figure 4.10). These all have a curved shape that starts and flattens around different values of time-based overlap. The fourth class goes down at the end, but there are fewer discussions that reach the length of 86 posts. Classes 1 and 2 have a slower increase in alignment than the other classes.

One discussion stands out in class 4 of bin 4, having a much lower alignment than the other discussions. Inspecting this discussion showed that it was in fact not a true discussion, but a movie ABC.

For bin 5 (Figure 4.12), the found classes again have the same curved shape of alignment with different start points and flattening points. The fourth class has a less steep alignment than the other three classes. The second class contains a discussion that stands out which starts off with a very low alignment and then reaches a very high alignment. Inspecting this discussion showed again that this conversation is in fact not a true discussion, but another continuation story game.

Another noticeable thing, in the third class, is that there are some downward spikes near the end from separate discussions. After investigating them, it showed that they were mostly true discussions, about religion, the Constitution, and slavery. These discussions contain some parts where the authors strongly oppose each other. One is not really a discussion but more a conversation where people even respond that the conversation is not going anywhere and they start arguing about that.

For bin 6 (Figure 4.13), the trends of the alignment are less smooth than the previous bins, as the number of discussions in this bin is lower. The shapes are quite similar, starting and flattening at different time-based overlaps, except for the fifth class which has a more flat alignment trend. The third graph is also different, in that the trend of alignment goes down at the end. This class also has a lower cardinality, which means that the trend is less smooth than for the other classes.

For bin 7 (Figure 4.15), we see the same outlier that was found before (see Section 4.4.2) in class 2. The other classes again have a more curved alignment trend and flatten at some level of time-based overlap. The third and the fifth classes seem to have a very similar shape, curved, starting a bit below an alignment of 0.4 and flattening around an alignment of 0.85-0.9.

For bin 8 (Figure 4.17), the clusters appear to be more separate, with the alignment trends having the same curved shape but starting and flattening all around different time-based overlaps. The fourth class has a lower cardinality and its alignment trend is therefore less smooth. This bin also has the lowest number of discussions among all bins, which shows in the graphs as they are less dense.

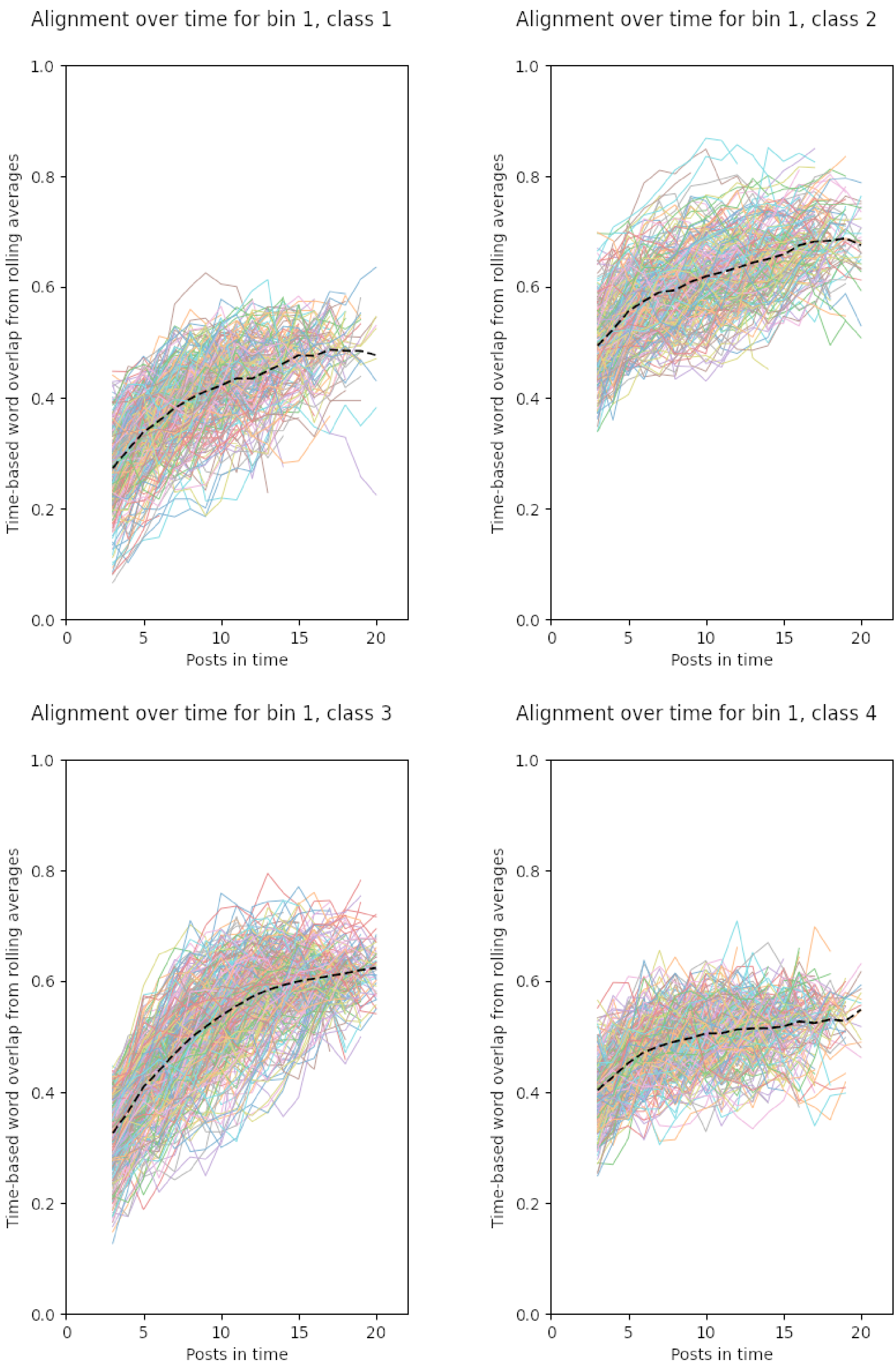


Figure 4.5: Best time series clusters for bin 1 (lengths 7-22)

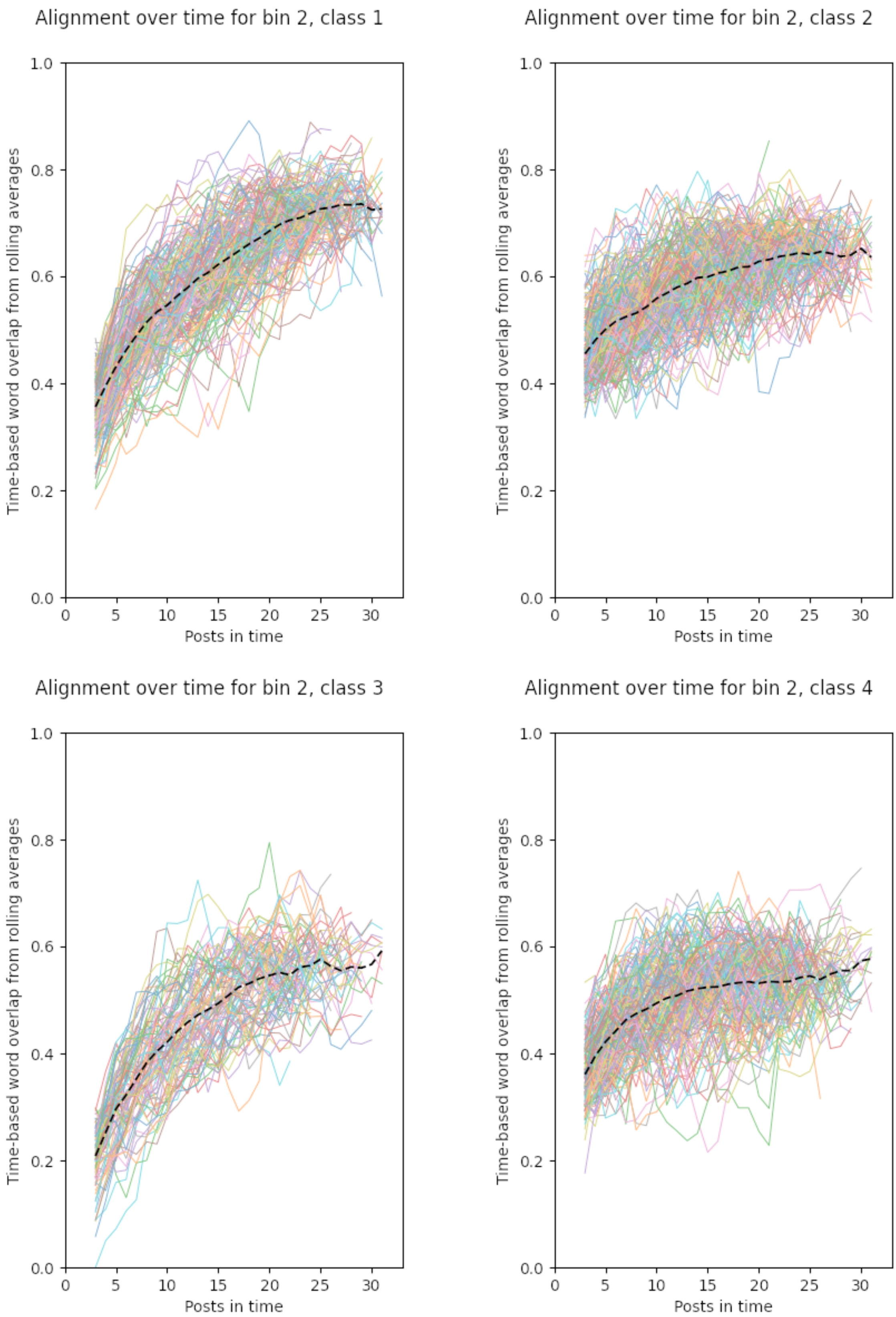


Figure 4.6: Best time series clusters for bin 2 (lengths 23-33), part 1

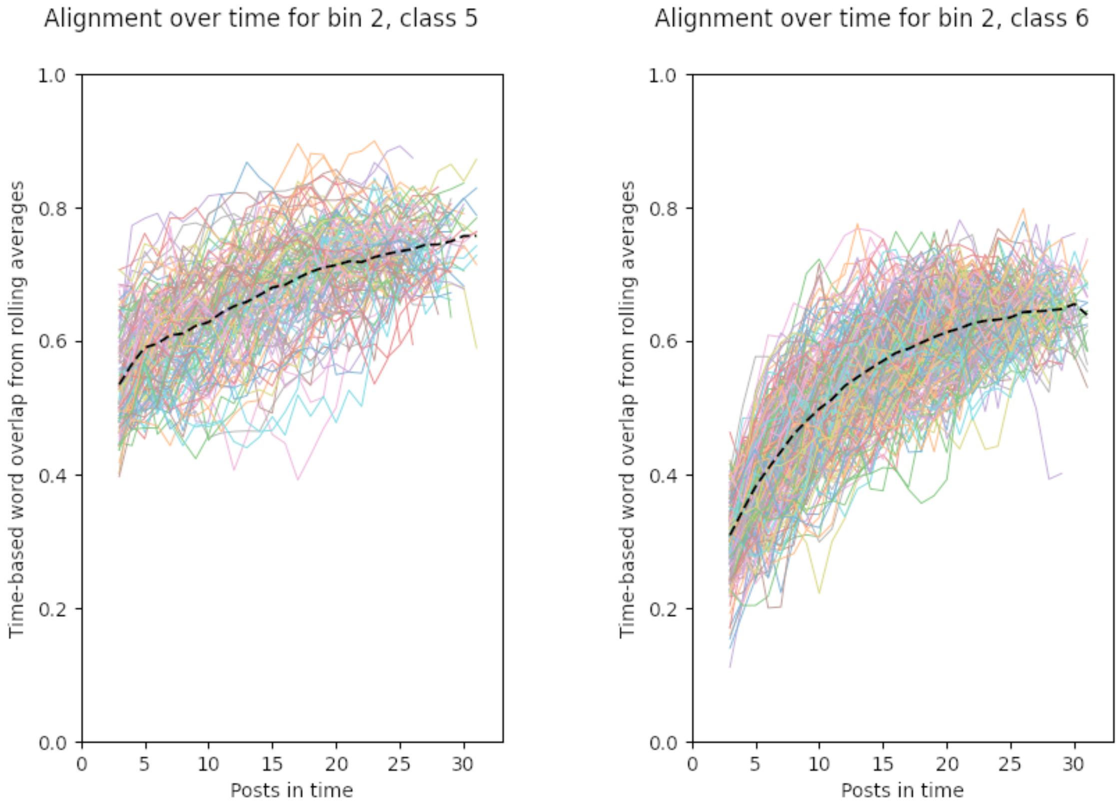


Figure 4.7: Best time series clusters for bin 2 (lengths 23-33), part 2

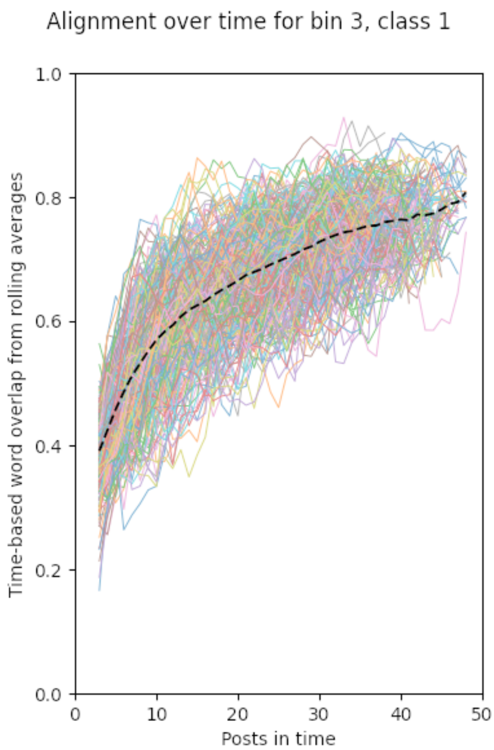


Figure 4.8: Best time series clusters for bin 3 (lengths 34-50), part 1

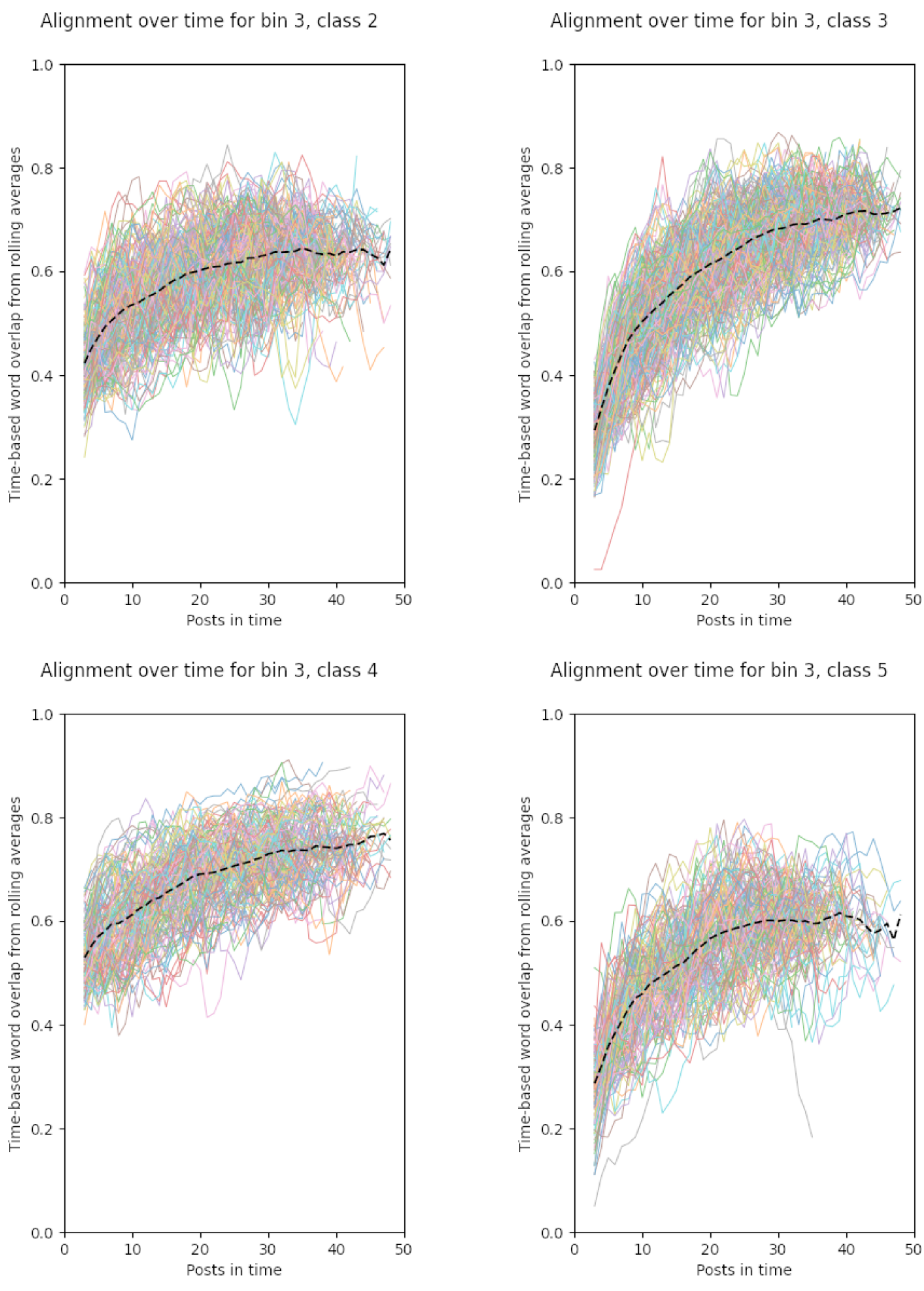


Figure 4.9: Best time series clusters for bin 3 (lengths 34-50), part 2

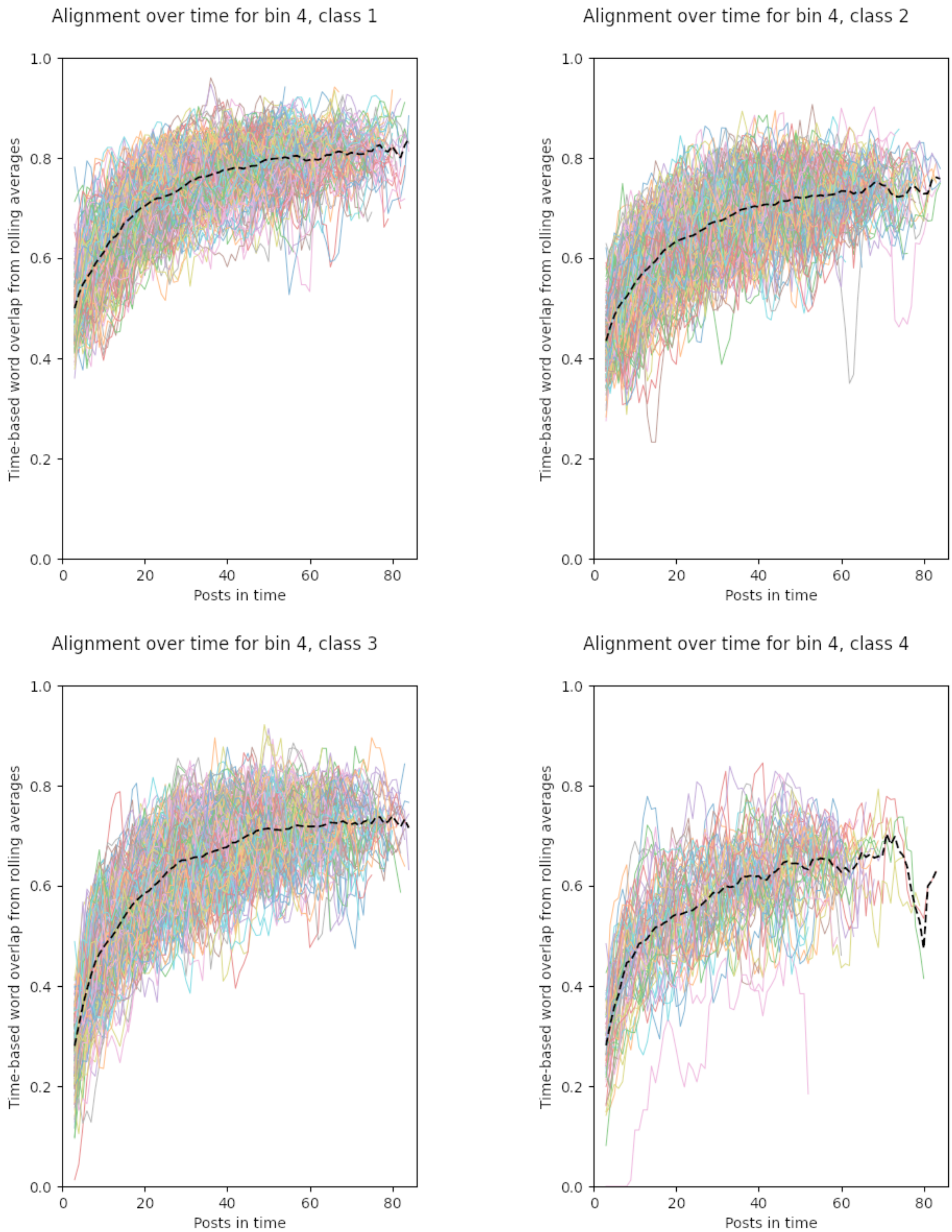


Figure 4.10: Best time series clusters for bin 4 (lengths 51-86), part 1

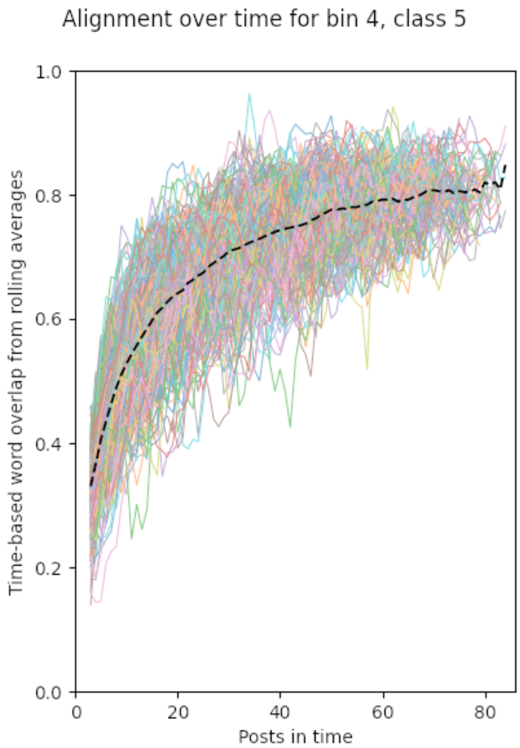


Figure 4.11: Best time series clusters for bin 4 (lengths 51-86), part 2

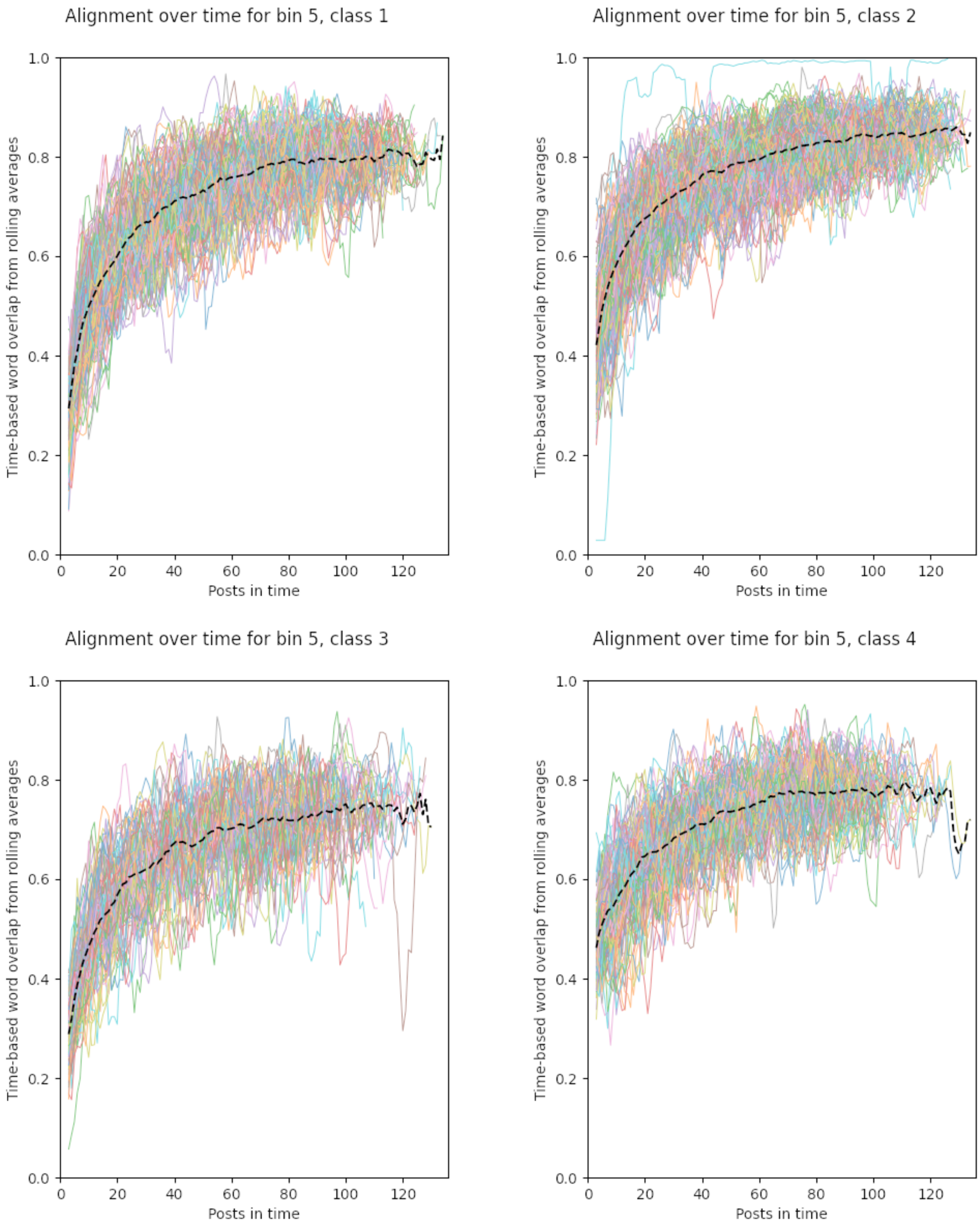


Figure 4.12: Best time series clusters for bin 5 (lengths 87-136)

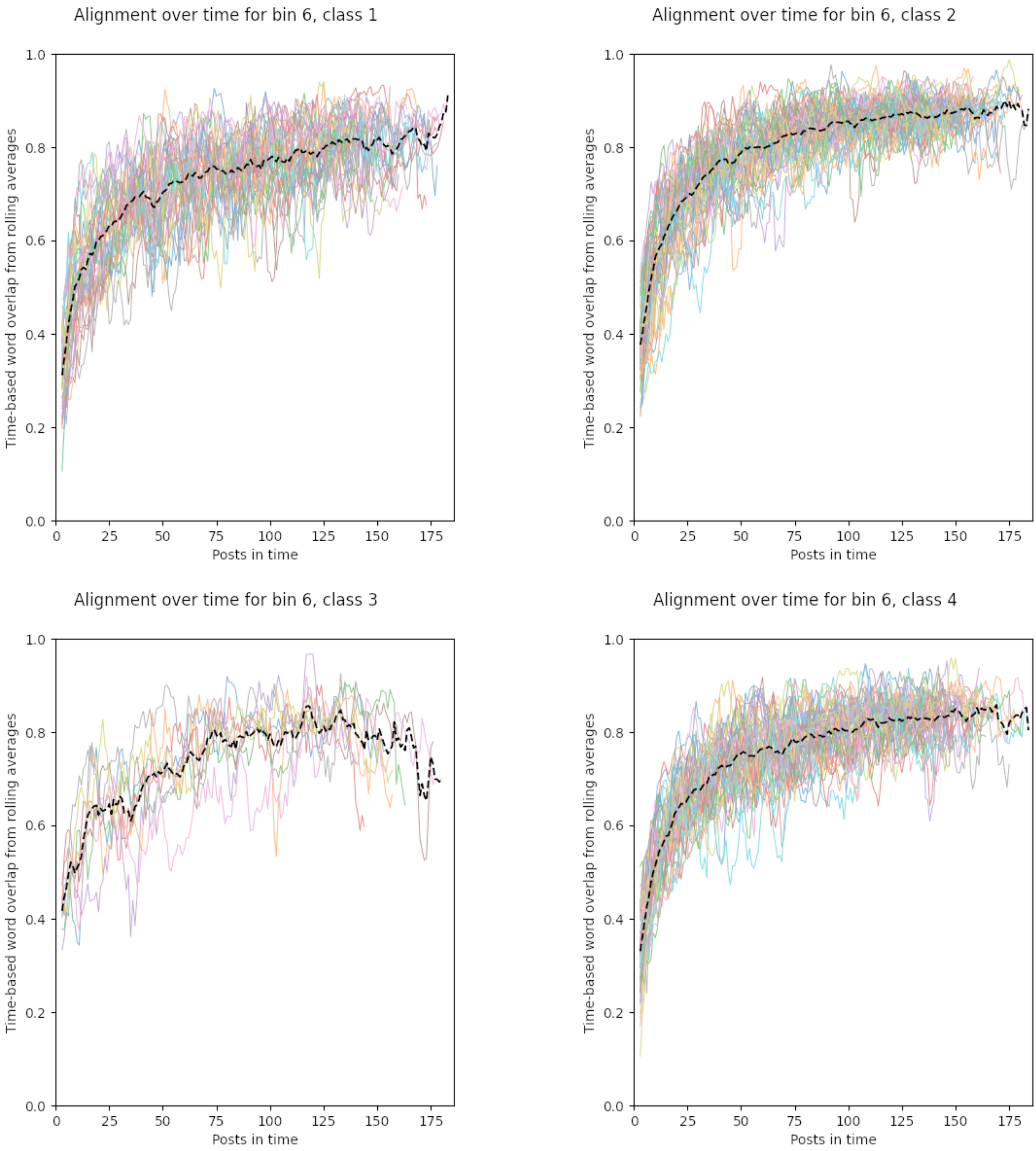


Figure 4.13: Best time series clusters for bin 6 (lengths 137-186), part 1

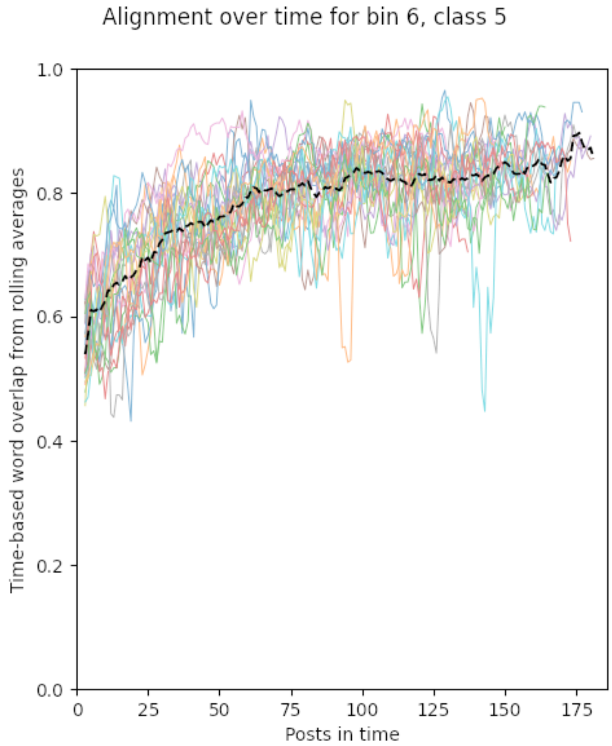


Figure 4.14: Best time series clusters for bin 6 (lengths 137-186), part 2

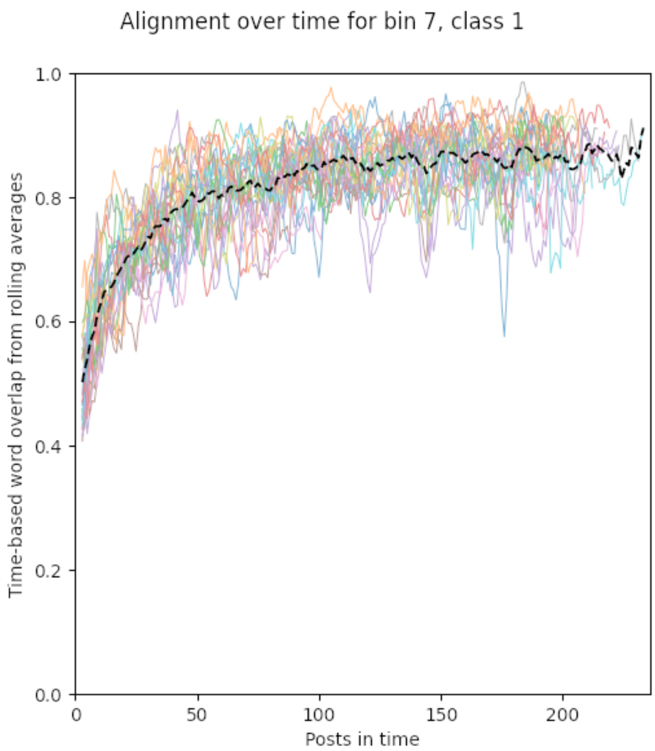


Figure 4.15: Best time series clusters for bin 7 (lengths 187-236), part 1

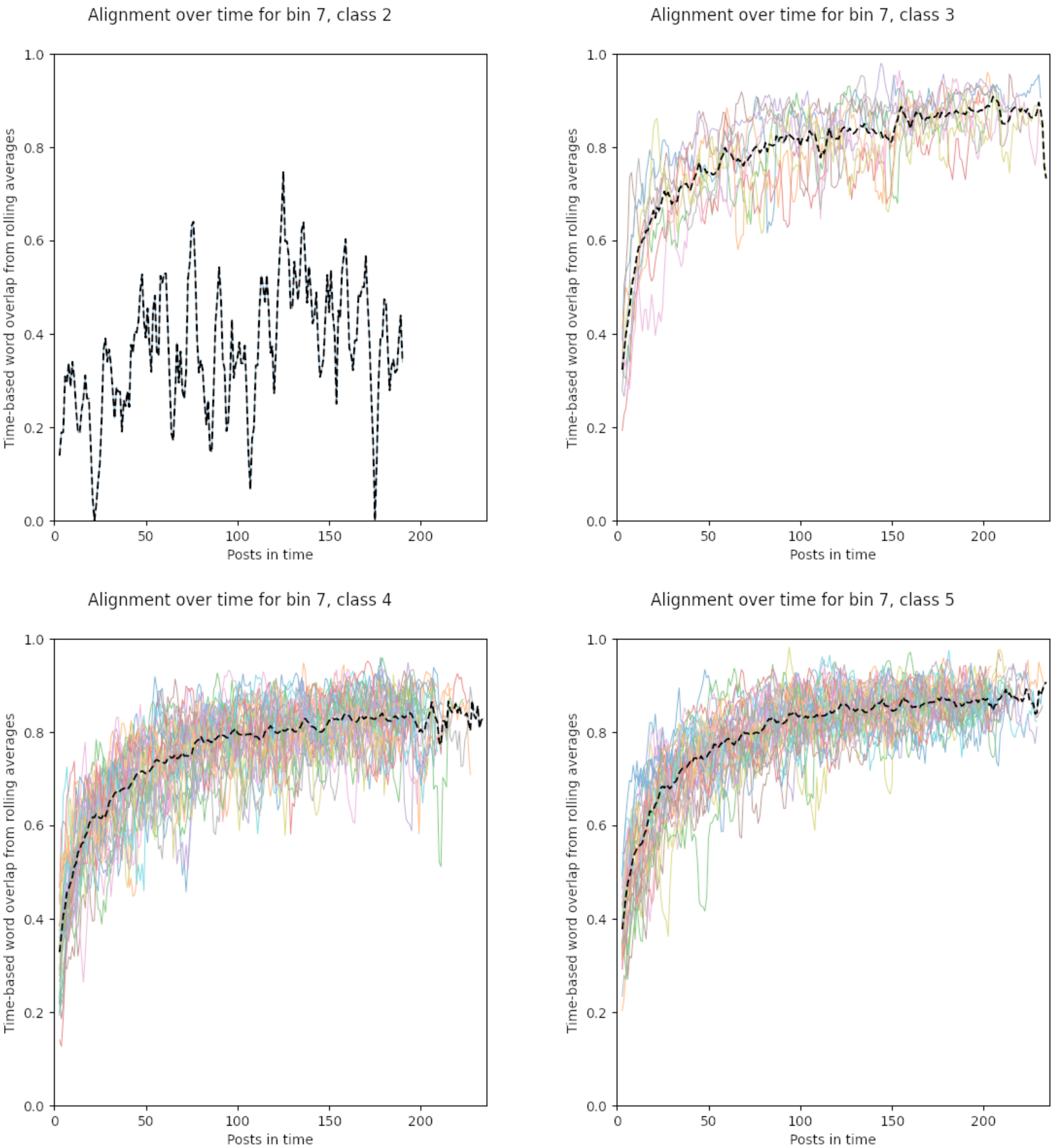


Figure 4.16: Best time series clusters for bin 7 (lengths 187-236), part 2

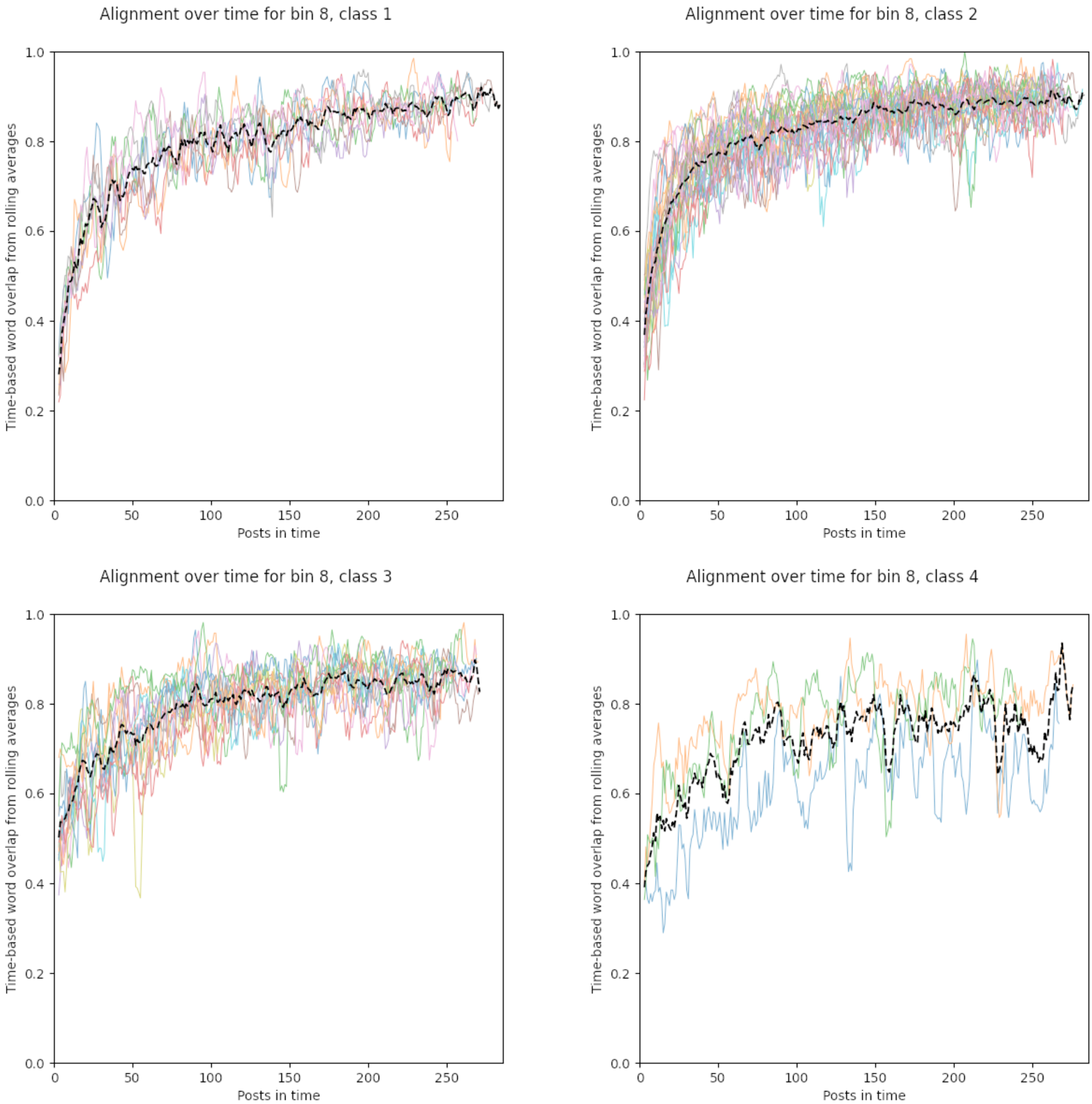


Figure 4.17: Best time series clusters for bin 8 (lengths 237-286)

4.5 Discussion

The results imply that people discussing online appear to reach a natural peak in average time-based overlap and that people participating in online discussion copy each other's words more and more towards the end of the discussions, where it reaches equilibrium around a time-based overlap of 0.6 for shorter discussions and up to 0.9 for longer discussions. The longer the discussions, the higher the time-based overlap seems to grow.

Longer discussions often have a steeper increase in alignment in the beginning, while shorter discussions have (more) classes where the increase is less curved and more linear. Alignment trends of shorter discussions have more different shapes than the longer trends.

Several things can be noted about this first analysis. The proposed time-based overlap method has a noticeable drawback. It computes overlap with vocabulary that was used before, excluding the vocabulary of the current author, which increases with every new post. Larger vocabularies make it more likely that a word in a post occurs in that vocabulary as opposed to a smaller vocabulary. Therefore, it is logical that using this measure over time, the time-based overlap increases. However, it was shown that priming over time decays, meaning that older terms have a lower influence on alignment than older terms. (Reitter et al., 2006). To get more representative results, some kind of cost could be added to vocabulary length or the distance in which the word previously occurred, to counteract this increasing effect. Future work could experiment with these suggestions. Nevertheless, the manner in which it currently increases or decreases is still very interesting, and patterns can still be extracted.

Another drawback is that the time-based overlap does not separate authors, it does not compute different time-series per author. Giles et al. state that authors do not necessarily align symmetrically (1991), so by combining all overlap, we disregard some potentially interesting information. Future work could investigate the time-based overlap per author per discussion and see if that results in different classes.

Furthermore, it could be that certain authors already inherently speak similarly, for instance, due to their backgrounds (homophily) (Doyle et al., 2016), which is not included in the time-based overlap. This could mean that the overlap presented in this study is higher than it actually is. Future work could experiment with including prior probabilities of authors using the word into the time-based overlap.

Another thing that should be noted is that currently, only the lexical (word) alignment level is inspected, though the other levels (word category, syntactical, and semantical level) might provide additional insights, or might possibly be more or less present than the currently inspected level. Future work could investigate the presence of other levels of alignment.

Unfortunately, there is no baseline yet of alignment in conversations or discussions, as previous works investigated alignment in relation to other subjects such as agreement (van der Pol et al., 2016), or per specific word category (Doyle & Frank, 2016). Furthermore, we introduced time-based overlap as a new measure, which means that it has not been used before. Therefore, it is not possible to say if this alignment we found is higher or lower than other situations, because we do not know the alignment of the other situations. For future work, it would therefore be very interesting to investigate alignment in other situations, like casual online conversations, or discussions on other platforms. It could also look into the correlations between time-based overlap and other measures, to investigate if they are comparable.

Another side note is that for each class in each bin, the end of the chart has fewer discussions than the start of the chart, which means that the trends could be less representative at those ends and should be looked at with some skepticism. For instance in bin 5, the ends of the trends appear to show interesting patterns but are based on only very few discussions.

Further improvements that could be made to this analysis is to use an improved stopword list, which contains the words that occur most in the dataset, also called topic words. Some words are necessary to discuss certain topics (for instance "abortion"), which induces some kind of "forced" alignment. To improve the analysis of alignment, such topic words could be removed. This could potentially show different patterns, changing the height of the alignment. However, as these topic words are currently included at every timestep in all discussions, it is presumed that this would mostly move the shapes downward (having an overall lower alignment) but not change the shapes of the alignment trends themselves. To create such a topic list, inspiration could for instance be drawn from Wen et al. (2014).

Another note is that we have used a moving average with a window of 5. This window size is a parameter that could be tweaked, we do not know if and how the resulting classes from the clustering would change if we had chosen another window. Future work could investigate the effects of this window size.

The clustering could also be improved by experimenting with different clustering methods and using different metrics than inertia. Inertia might not be the best metric for evaluating clustering quality, as it only inspects the inter-class variance, not between the classes. However, as the alignment in discussions shows so much variance, it would seem impossible to find separate clusters that do not overlap.

A last remark on the time series analysis is how we treat *time*. Time in discussions is measured in terms of posts, not the actual time of posting, to simplify the problem. The time between posts might vary, which might in turn affect the results. Furthermore, as we have this notion of time, it is not clear how to best align conversations of different lengths in time. Currently, they are aligned at the beginnings of discussions (post zero), while the end is sparsely populated. We could just as well have aligned the ends, which might give different results. In future work, these other notions of time and their effects on the results could be researched.

4.6 Conclusion

The time-based overlap was applied to all messages in all discussions, and plotted in histograms, for all messages and for the average alignment per discussion, next to computing percentiles. With these results, we can now answer **Q1.1**. Lexical word alignment (based on the time-based overlap) is present in the discussions with a mean of 0.6 in online discussions on 4forums. The distribution of average alignment is bell-shaped and ranges between 0.2 and 0.9.

Next, discussions were divided into bins based on the number of posts. With the rolling average of the alignment, each bin was clustered to find trends of alignment. With that, we can now also answer **Q1.2**. Alignment in most classes in most bins increases with a curved shape, stabilizing at some level of time-based overlap. In a few classes (for instance bin 3) the trend goes down near the end. Trends mostly differ in at which alignment they start, how steep the increase is, and how high the alignment reaches.

Chapter 5

Sentiment analysis

This chapter discusses the second part of the thesis, which investigates the sentiment expressed in discussions. It answers the following subquestions:

Q2.1 What is the distribution of sentiment expressed by interlocutors during the discussions?
Q2.2 How does the expressed sentiment change over posts in the discussions?

5.1 Preprocessing for sentiment

After the general preprocessing as described in Section 3.3, one additional preprocessing step was applied for the sentiment analysis. To apply VADER, see the next section, the posts should be divided into sentences. To split posts into sentences, the `sent_tokenize` function of `NLTK.tokenize` was used.

Characters were not converted into lowercase, as all-caps words can indicate a higher arousal for a sentiment, and are informal intensifiers (Kouloumpis, Wilson, & Moore, 2011). Furthermore, stopwords were not removed, as that often does not improve performance in sentiment analysis (Jurafsky & Martin, 2023).

5.2 Sentiment metric

VADER is a simple rule-based model for general sentiment analysis (Hutto & Gilbert, 2014). It works on the sentence level and returns the proportions of the sentence that belong to a negative, neutral, and positive sentiment, and an overall compound score of the sentence. We are looking at the compound score, which ranges from -1 to 1, where -1 means that the overall sentiment expressed in the sentence is intensely negative, and 1 means intensely positive. A compound score of 0 means that that sentence's overall sentiment is neutral, not positive or negative.

VADER works based on a lexicon and rules (Hutto & Gilbert, 2014). Hutto and Gilbert created a new sentiment lexicon specifically attuned to microblog-like contexts, which fits the 4forums dataset. Features from such contexts include emoticons (such as ":-)"), sentiment-related acronyms (e.g. "LOL" and "WTF"), and slang (e.g. "meh" and "giggly"). The lexicon was evaluated by human evaluators. For each word, a sentiment score is extracted from this lexicon. VADER then combines the scores from this lexicon with five rules that address sentiment intensity in sentences:

1. Punctuation increases the magnitude of sentiment intensity in the sentence without modifying the semantic orientation. E.g."The food here is good!!!" conveys more intensity than "The food here is good".
2. Capitalization increases the magnitude of the sentiment intensity of the capitalized word without affecting the semantic orientation. E.g. "The food here is GREAT" conveys more intensity than "The food here is great".
3. Degree modifiers (such as "extremely" and "marginally") impact the sentiment intensity of the subsequent word by either increasing or decreasing its intensity. E.g. in "The food here is really good" is more intensely "good" than in "The food here is good", but "The food here is marginally good" conveys a less intense sentiment in "good".
4. The contrastive conjunction "but" indicates a shift in sentiment polarity, with the sentiment of the text following the conjunction in the sentence being dominant (increasing the text after the conjunction with 50% intensity, while decreasing the text before the conjunction with 50%). E.g. "The food here is great, but the service is terrible" conveys a negative sentiment as the second part dictates the overall rating.
5. Negation flips the polarity of the part of the sentence belonging to the negation. E.g. "The food here isn't good" is negated and conveys a negative sentiment.

When the rules are applied to the scores extracted from the lexicon, VADER returns the proportions of the sentence that belong to the negative, neutral, and positive sentiment, and provides the compound score. The compound score is a normalized, weighted composite score. The authors in the original paper use classification thresholds for the compound score at -0.05 and 0.05 (Hutto & Gilbert, 2014), though the documentation suggests using thresholds at -0.5 and 0.5[1]. The documentation of VADER states texts longer than one sentence can be analyzed by averaging the sentiment scores of individual sentences.

VADER was chosen as a measure for the sentiment analysis as it was found to rank the best for three-class sentiment analysis (Ribeiro et al., 2016), it was created specifically for extracting sentiments in social media (Hutto & Gilbert, 2014), and it fits the dataset. Another perk is that it is easy to use, as it is available as a Python tool.[2]

[1]Documentation of VADER about the thresholds: https://vadersentiment.readthedocs.io/en/latest/pages/about_the_scoring.html

[2]VADER tool: https://vadersentiment.readthedocs.io/en/latest/index.html

5.3 Methods

A notebook with the code for the methods that follow can be found at the Github repo[3], see step 3: sentiment analysis in the README.

5.3.1 Distribution of sentiment (Q2.1)

To find the distribution of sentiment in the discussions, the `compound` polarity score of the VADER analyzer (`SentimentIntensityAnalyzer`) of the `vaderSentiment` package was computed for all preprocessed messages in the discussions. The distribution of the sentiment scores of all messages was plotted in a histogram, and the mean, min, max, and 1st, 5th, 10th, 50th, 90th, 95th, and 99th percentiles were computed.

Furthermore, the average was taken of the sentiment scores per discussion, a histogram of the distribution was plotted and the same statistics as previously were computed. The same was done for the minimum and maximum sentiment scores per discussion, and in addition, a joint distribution was plotted for this minimum and maximum. This information about the sentiment distribution gives a first insight into which sentiment is expressed in our dataset.

5.3.2 Sentiment changing over posts (Q2.2)

Using the computed sentiment scores, we first plotted the sentiment of all posts for all discussions. A rolling average with a window of 5 was applied to smoothen the time series and to be able to find the general trends rather than the exact trends. We disregard the first and the last time points where the window of 5 cannot be applied. The rolling average of the sentiment scores was plotted.

Then, time series clustering was again applied to find trends in our sequential data with `TimeSeriesKMeans` from the `tslearn.clustering` package. For consistency, we again use the DTW metric for cluster assignment. We also split up the discussions into the same bins as were created for analyzing the alignment for consistency (see Table 4.3).

The number of clusters, k, was again fine-tuned for each bin by running the `TimeSeriesK-Means` with k in the range $[1, 10]$ and fitting it on the sentiment scores for the discussions in that bin. Similar to the alignment analysis, the optimal k was chosen by using the elbow method of the inertia per k. The classes for all k for each bin were also plotted in line charts to support choosing the optimal k. Figure 5.1 can be used as a reminder of what is meant by *bin* and *sentiment class*.

To get the optimal clustering, we run the clustering for each bin with their optimal k five times, and store the model with the lowest (best) inertia to obtain the most optimal clustering per discussion length bin. For each discussion length bin, the optimal clustering was plotted per class, showing the different trends of sentiment over time (posts) per discussion length.

[3] `https://github.com/SuzannaWentzel/Sentiment-Alignment-Interplay`

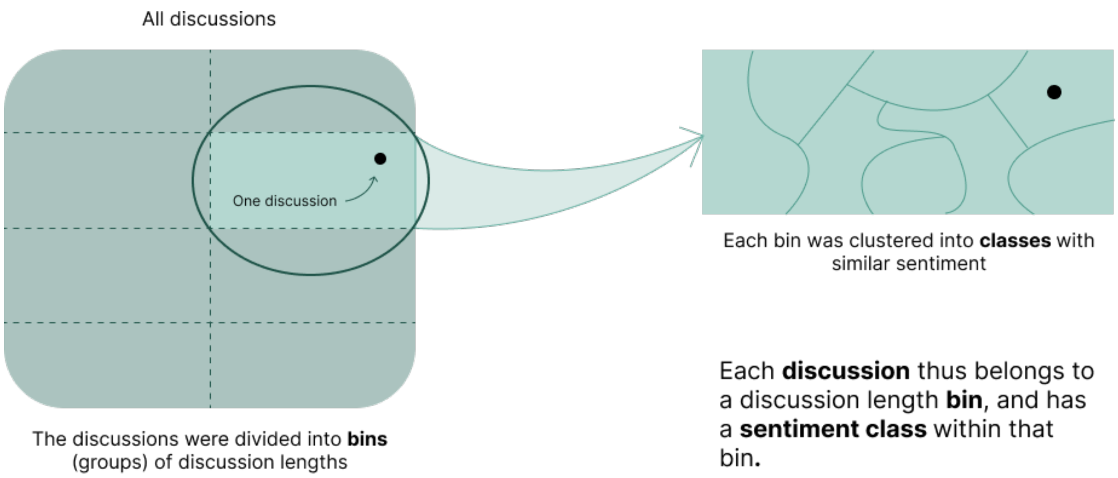


Figure 5.1: What is meant by bins and sentiment classes

5.4 Results

5.4.1 Distribution of sentiment in discussions

The statistics of all sentiment scores of posts are shown in Table 5.1. The accompanying histogram is shown in Figure 5.2. These graphs show the number of posts per sentiment scores, whereas the second graph is a log-scaled version of the first, such that the lower number of posts per sentiment does not disappear. The third graph shows a zoomed-in version of the first, cutting off the high peak at 0.0.

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	**90th**	**95th**	**99th**	**Max**
sentiment score	0.0115	-0.9975	0.-0.7052	-0.4180	-0.2902	0.0000	0.3072	0.4236	0.6875	0.9943

Table 5.1: Statistics of the sentiment score of all posts

The distribution of the sentiment scores of all posts is bell-shaped, with a mean of 0.01. The scores spread out between the possible ends of the sentiment score, which shows a large variety of expressed sentiments within the dataset.

The statistics of the average sentiment scores in discussions are shown in Table 5.2. The accompanying histogram is shown in Figure 5.3. The graphs in the figure show the number of discussions per average sentiment score. The second graph is a log-scaled version of the first, such that the lower number of discussions per sentiment score does not disappear.

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	**90th**	**95th**	**99th**	**Max**
sentiment score	0.0132	-0.4017	-0.2276	-0.1404	-0.1005	0.0212	0.1133	0.1461	0.2093	0.3546

Table 5.2: Statistics of the average sentiment scores of all discussions

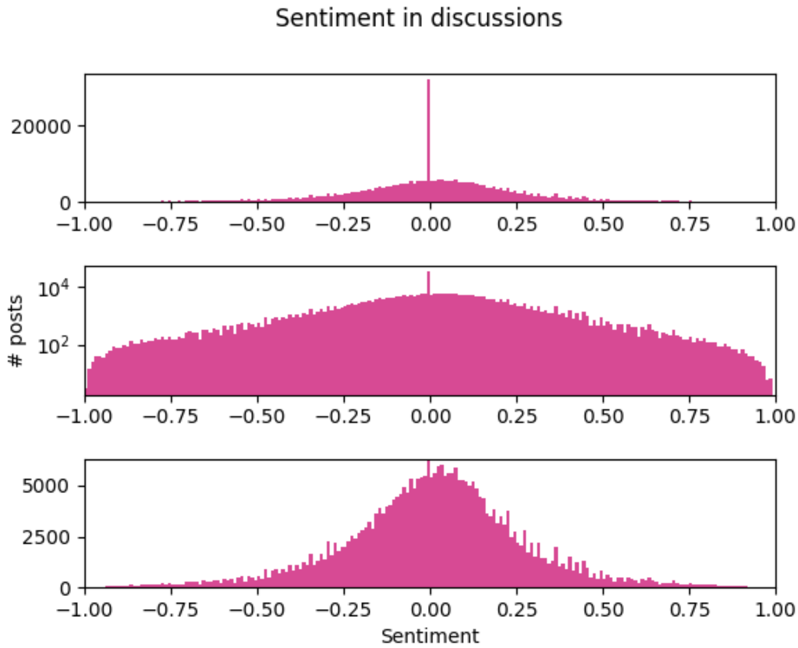


Figure 5.2: Distribution of sentiment scores of posts

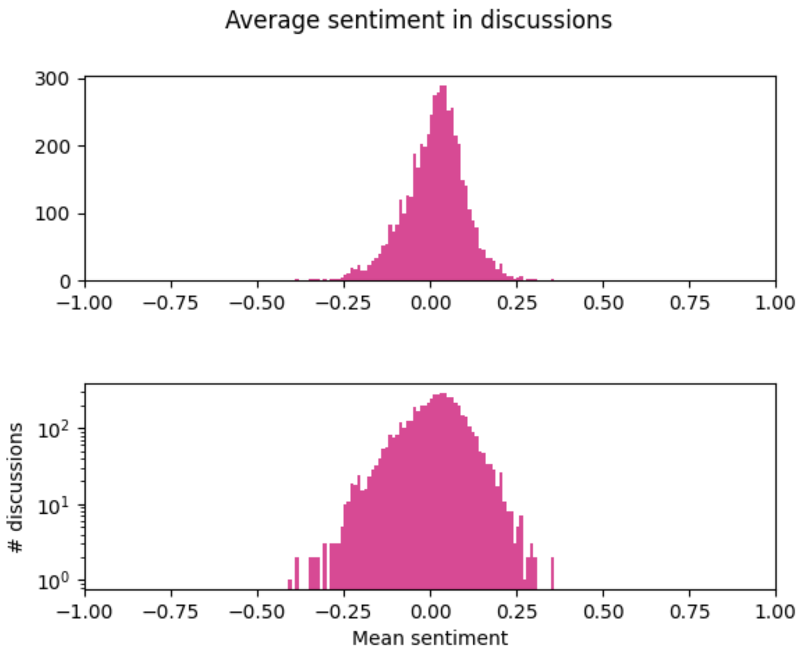


Figure 5.3: Distribution of average sentiment scores in discussions

The distribution of the mean sentiment in discussions is also bell-shaped, again with an average of 0.01. The range of average sentiments is much smaller than the range of the sentiments of all posts. This makes sense as we look at it from a higher level; the extremes are balanced out.

The following excerpt is taken from a discussion (about abortion) with a lower average sentiment score (of -0.2, from the 1st percentile):

Example 5.4.1: Discussion with lower sentiment

...
Message 21 (author 22): no matter it is legal or illegal to our laws, abortion is killing a life, so it is people's crime !!! emoticonXMad (Sentiment score: **-0.30**)
Message 22 (author 3): "Technically it is not a crime, because it is legal. If it were illegal then it would be a crime." Carman This isnt under crime debates cuz its technically a crime its just that it should be a crime its killing and you know it. (Sentiment score: **-0.38**)
Message 23 (author 570): Yes, I know it's a killing... I never said that it wasn't, I only said that LEGALLY it is not considered to be a crime (at this point).Ok, I thought you were asking from the legal standpoint... sorry... (Sentiment score: **-0.35**)
Message 25 (author 642): What the Hell kind of a question is that? (Sentiment score: **-0.68**)
Message 26 (author 38): Abortion isn't a legal crime. How ever, in all other ways, it certainly is a crime. I think the moderator should move this to abortion discussion. (Sentiment score: **0.03**)
...

The following excerpt is from a discussion (about the Chinese stock market) with a characteristic sentiment score (0.01, mean sentiment):

Example 5.4.2: Discussion with characteristic sentiment

...
Message 12 (author 127): Limited working weeks enhance productivity. Britain opts out of the working time directive (max 48 hour week) and the result is low producivity. If people work shorter hours then, up to a point, they will do pretty much the same work, but in less time. (Sentiment score: **-0.08**)
Message 13 (author 1077): You really take the kool-aid, don't you? (Sentiment score: **0**)
Message 14 (author 317): Heh, I'm sorry but I was too busy shorting US semiconductor stocks for the past year. :p (Sentiment score: **0.07**)
Message 15 (author 1077): Like our stock market is not an leading indicator of future economic activity? So the Chinese stock market is different, is it? They have repealed the laws of economics? Really, why don't you tell us all about our market and "parity with valuation?" (Sentiment score: **0.09**)
...

This last excerpt is taken from a discussion (about music styles and artists) with a higher average sentiment score (of 0.2, from the 99th percentile):

Example 5.4.3: Discussion with higher sentiment

...

Message 12 (author 343): Of course the blues was an influence. The blues was an influence on all rock & roll, not excluding Buddy Holly, and every rock guitarist owes a debt of graditude to Chuck Berry. But, like someone said in Rolling Stone "Listen to the songs on the first three Beatles albums. Take their voices off, and it's Buddy Holly." (Sentiment score: **0.03**)

Message 13 (author 1017): My favorite band is Pink Floyd, but I love the Eagles and a lot of other classic rock, as well as a lot of the good country, before it became so much like pop (Sentiment score: **0.96**)

Message 14 (author 129): I was never a big fan of the Beatles or Buddy Holly, they were pretty much mainstream sounds. Liked a few of their songs but for rock in general preferred Cream, Buffalo Springfield, Pink Floyd, etc. and many others and the groups/artists that evolved from those beginnings. (Sentiment score: **0.18**)

Message 15 (author 456): Yeah, I also have to say my favorite band is Pink Floyd. Wish You Were Here is such a good song. For best vocalist, personally, I'd have to give that to Art Garfunkel of Simon and Garfunkel. Bridge over Troubled Water.....wow.... (Sentiment score: **0.37**)

...

The statistics of the minimum sentiment scores in discussions are shown in Table 5.3. The accompanying histogram is shown horizontally in Figure 5.4. The distribution of the minimum sentiment scores of discussions is bell-shaped but skewed to the right. The distribution has a mean of -0.58 and a median of -0.60.

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	**90th**	**95th**	**99th**	**Max**
sentiment score	-0.5826	-0.9975	-0.9675	-0.9139	-0.8720	-0.5994	-0.2695	-0.1936	-0.0746	0.0206

Table 5.3: Statistics of the minimum sentiment scores of all discussions

The statistics of the maximum sentiment scores in discussions are shown in Table 5.4. The accompanying histogram is shown vertically in Figure 5.4. The distribution of the maximum sentiment scores of discussions is bell-shaped but also skewed to the right. The distribution has a mean of 0.61 and a median of 0.62.

	Mean	**Min**	**1st**	**5th**	**10th**	**50th**	90th	**95th**	**99th**	**Max**
sentiment score	0.6053	-0.0562	0.1598	0.2786	0.3482	0.6240	0.8555	0.8990	0.9529	0.9943

Table 5.4: Statistics of the maximum sentiment scores of all discussions

From the distributions of the minimum and maximum sentiment scores per discussion, we can indeed see that the extremes have been balanced out in the average sentiment score distribution, as the distributions for the minimum and maximum sentiment scores indeed have a much bigger range than the average sentiment score.

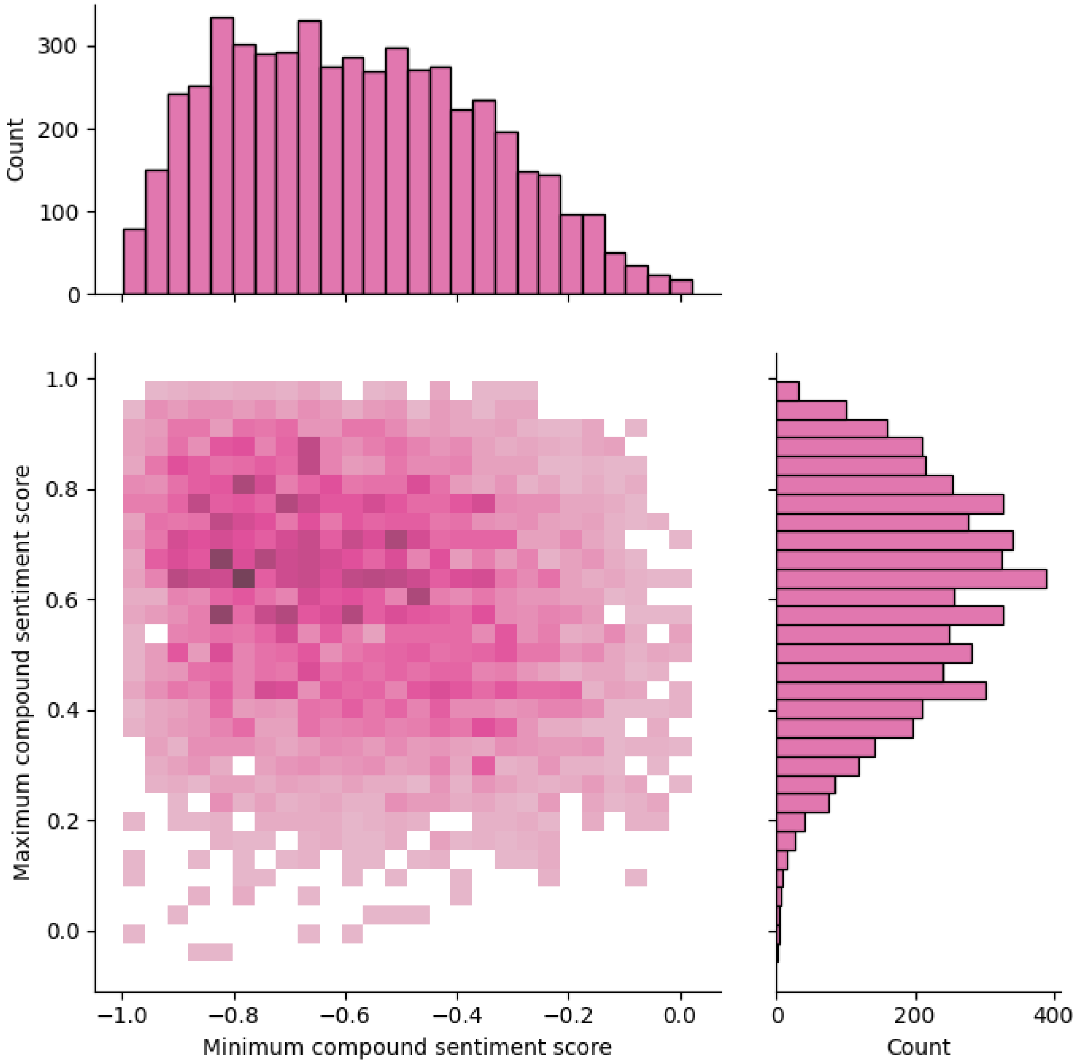


Figure 5.4: Joint distribution of minimum and maximum sentiment score per discussion

5.4.2 Finetuning clustering

The plots for the sentiment scores changing over posts and the rolling averages can be found in Appendix D. The plots for finding the optimal cluster sizes per bin can be found in Appendix E.1. They plot the inertia per amount of clusters. The optimal amount of classes per discussion length bin derived from these figures is shown in Table 5.5.

Bin	**Discussion length**	**# Discussions**	**Optimal k**
1	7-22 posts	1089	6
2	23-33 posts	1049	5
3	34-50 posts	1058	7
4	51-86 posts	1052	6
5	87-136 posts	556	8
6	137-186 posts	174	6
7	187-236 posts	106	7
8	237-286 posts	52	6

Table 5.5: Optimal k per discussion length bin

Some of the elbows show some interesting behavior, for instance, the peaks in Figures E.5 (bin 5) and E.7 (bin 7), and the plateau in Figure E.8 (bin 8).

Looking into the plotted classes of bin 5 (see Appendix E.2), we see that with $k = 9$, there are multiple classes with lower cardinalities. This could cause the other classes to have higher cardinalities, with potentially higher variances, causing the peak. The clustering results with lower k's all have classes with relatively similar cardinalities. Since $k = 8$ has the lowest inertia, we go for $k = 8$ for bin 5.

Looking at the classes of sentiment trends found in bin 7 (see Appendix E.3), we see that classes have lower cardinalities, as this bin contains fewer discussions. Furthermore, the trends of the sentiment score all appear to be rather flat. $k = 7$ was chosen as $k = 5$ and $k = 6$ have higher inertia and classes with more overlap, and $k = 8$ has two classes with very low cardinalities.

Looking at the plotted classes for bin 8 (see Appendix E.4), the clusterings with $k = 7$ and $k = 8$ have multiple classes with low cardinalities, and $k = 5$ has a lot of overlap between the trends of the classes. So, $k = 6$ was chosen.

5.4.3 Time series clustering on sentiment scores

The best performing time series clustering models for each bin are shown in Figures 5.5 - 5.19. The figures show the discussions per class that the clustering has found over time (posts) and the average trend of the sentiment score in that class. The figures differ in the number of classes, see Section 5.4.2.

For bin 1 (Figures 5.5 & 5.6), the six classes are different in shape and/or height. Classes 1, 4, and 6 show a decrease where the sentiment becomes less positive, though their shapes differ. The sentiment of the first class first gets more positive before moving to the negative polarity. The fourth class has a steep decrease from a more positive to a neutral sentiment, whereas the sixth class gradually moves towards a more negative sentiment. Class 2 and 3 increase respectively from a slightly negative and a slightly neutral sentiment towards a slightly more positive sentiment, though the increase appears to be minimal. Class 5 has a lower cardinality but shows an increase from a lower sentiment towards a more neutral sentiment, in a more curved shape.

For bin 2 (Figures 5.7 & 5.8), again separate classes were found. Classes 1 and 2 have a trend of increasing sentiment scores over posts. Class 1 starts around a neutral sentiment and increases to a slightly positive sentiment, whereas class 2 starts at a slightly negative sentiment and rapidly increases to a more neutral sentiment. Class 4 and 5 have decreasing sentiment scores, with class 4 going from neutral to slightly negative and class 5 from positive to neutral. Class 3 is also interesting, with the sentiment score slightly going down from a more positive sentiment and then up again after around 17 posts.

Bin 3 (Figures 5.9 & 5.10) has some more overlap between classes than the previous bins and is more flat and neutral. Classes 1, 2, and 6 decrease in sentiment. Classes 1 and 6 go from slightly positive to more neutral, and class 2 from neutral to slightly negative. The sentiment trend of class 4 stays flat around a neutral sentiment. Class 7 has the lowest trend of sentiment scores out of all classes, which stays approximately flat as well. This class has a lower cardinality than the other classes. Class 5 increases from a negative sentiment to a neutral sentiment. The trend of class 3 starts with a slightly negative sentiment increases to neutral, decreases again, and increases to a slightly positive sentiment.

Bin 4 (Figures 5.11 & 5.12) also has a lot of overlap between the classes of sentiment trends. It has two classes, class 2 and class 6, where sentiment goes downward towards a neutral sentiment, where class 6 starts more positively than class 2. Classes 3 and 5 have an increasing sentiment score. Class 3 starts neutral and increases to a more positive sentiment. Class 5 starts slightly negative and increases to a neutral sentiment. Class 4 starts with a slightly negative sentiment, increases to neutral, and decreases again after around 30 posts. Class 1 shifts around a neutral sentiment. Classes 1 and 4 have a lower cardinality than the other classes of bin 4, which is why their trends are also less smooth.

The sentiment trends of bin 5 (Figures 5.13 & 5.14) also overlap. There is mostly some variance at the beginning of the shapes, after which they reach a plateau. Class 1 has the overall highest sentiment trend, though it starts more positive and decreases slightly towards a neutral sentiment. Classes 2, 4, and 8 stay flat at a neutral sentiment, with slight increases and decreases. Class 3 has the overall lowest sentiment, and starts more negative and increases slightly towards a more neutral sentiment. Class 5 starts more positively and flattens at a neutral sentiment. Class 6 increases most of all classes, going from a slightly negative to a slightly positive sentiment. The sentiment score trend of class 7 is flat but is not exactly neutral and tends slightly more toward a positive sentiment. Starting from discussion lengths bin 5, the bins have decreasing numbers of discussions, which means that the cardinalities per class are automatically lower than the previous discussion lengths bins.

This shows in the trends, as they are less smooth than in other bins.

Bin 6 (Figures 5.15 & 5.16) also has a lot of overlap between its classes. The trends are less smooth because of the lower cardinality per class. Apart from the zig-zagging, the trends of classes 2, 3, 4, and 5 are rather flat, around an approximately neutral sentiment. Class 1 starts with a more negative sentiment and increases towards a more positive sentiment, similar to class 6.

Bin 7 (Figures 5.17 & 5.18) also has a lot of overlap and even lower cardinalities than the previous classes. Apart from the zig-zagging, the sentiment score trends are rather flat, with some slight changes in shapes. Class 1 goes from slightly positive to neutral to positive, class 2 goes down to slightly negative, class 3 goes down from slightly positive to slightly negative, to slightly positive to neutral, class 4 from neutral to slightly positive to negative, class 5 stays flat around slightly negative, class 6 goes from slightly positive to more neutral and class 7 stays flat around neutral.

Bin 8 (Figures 5.19 & 5.20) has even lower cardinalities in classes than bin 7, and class 5 has a particularly low cardinality (4 discussions). Again, apart from the zigzagging, the lines stay rather flat. Class 5 also stands out in having the lowest sentiment scores from all classes, around slightly negative.

Some trends are visible across bins, this is shown in Table 5.6. Most bins (6 out of 8) contain a class that has a trend around neutral, from neutral to positive, and from negative to neutral. 5 out of 8 bins have a class around positive, and from positive to neutral. Rare classes that only appear in one bin are from positive to more positive, from negative to more negative, and from positive to negative. Another remark is that at first glance, patterns growing more positive (higher in the table) seem to be more represented by classes than patterns growing more negative (lower in the table). However, one should note that this doesn't say anything about the number of discussions in those classes, as discussions are not evenly distributed over the classes, nor over the higher bins.

	Bin 1	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 6**	**Bin 7**	**Bin 8**
Positive - More positive	Class 3							
Neutral - Positive	Class 2	Class 1		Class 3	Classes 2, 4	Class 2		Classes 2, 4
Negative - Positive			Class 3	Class 1	Class 6			
Negative - Neutral	Class 5	Class 2	Class 5	Class 5		Class 1, 6	Class 7	
Negative - Less negative					Class 3			Class 5
Around positive		Class 3			Class 1	Class 4	Class 1	Class 1
Around neutral	Class 1		Class 4		Class 7	Class 3, 5	Classes 2, 4	Class 3
Around negative			Class 7	Class 4			Class 5	
Positive - Less positive			Class 6	Class 6			Class 6	
Positive - Neutral	Class 4	Class 5	Class 1	Class 2	Class 5			
Positive - Negative							Class 3	
Neutral - Negative		Class 4	Class 2		Class 8			Class 6
Negative - More negative	Class 6							

Table 5.6: Similar sentiment trends in bins

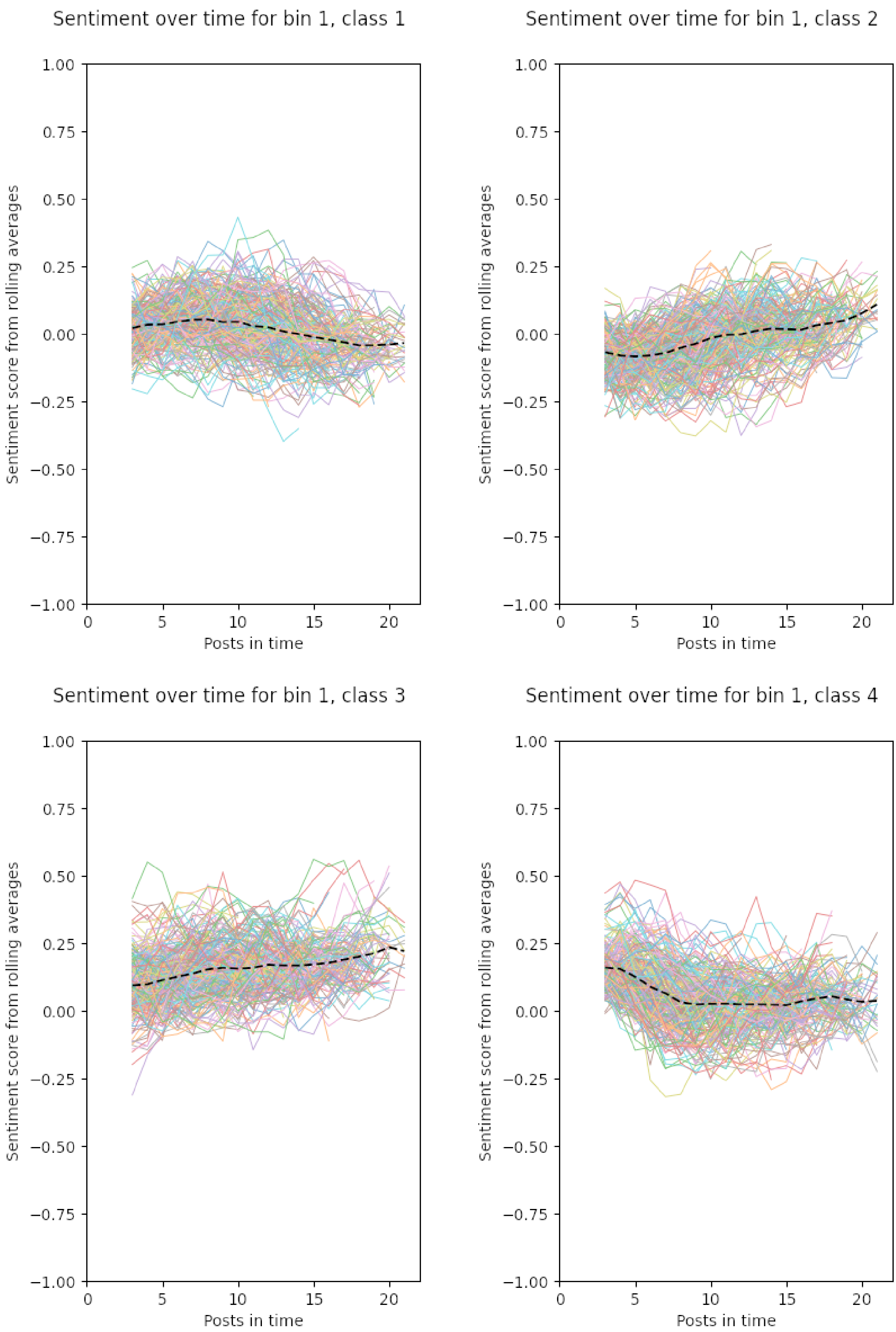


Figure 5.5: Best time series clusters for bin 1 (lengths 7-22), part 1

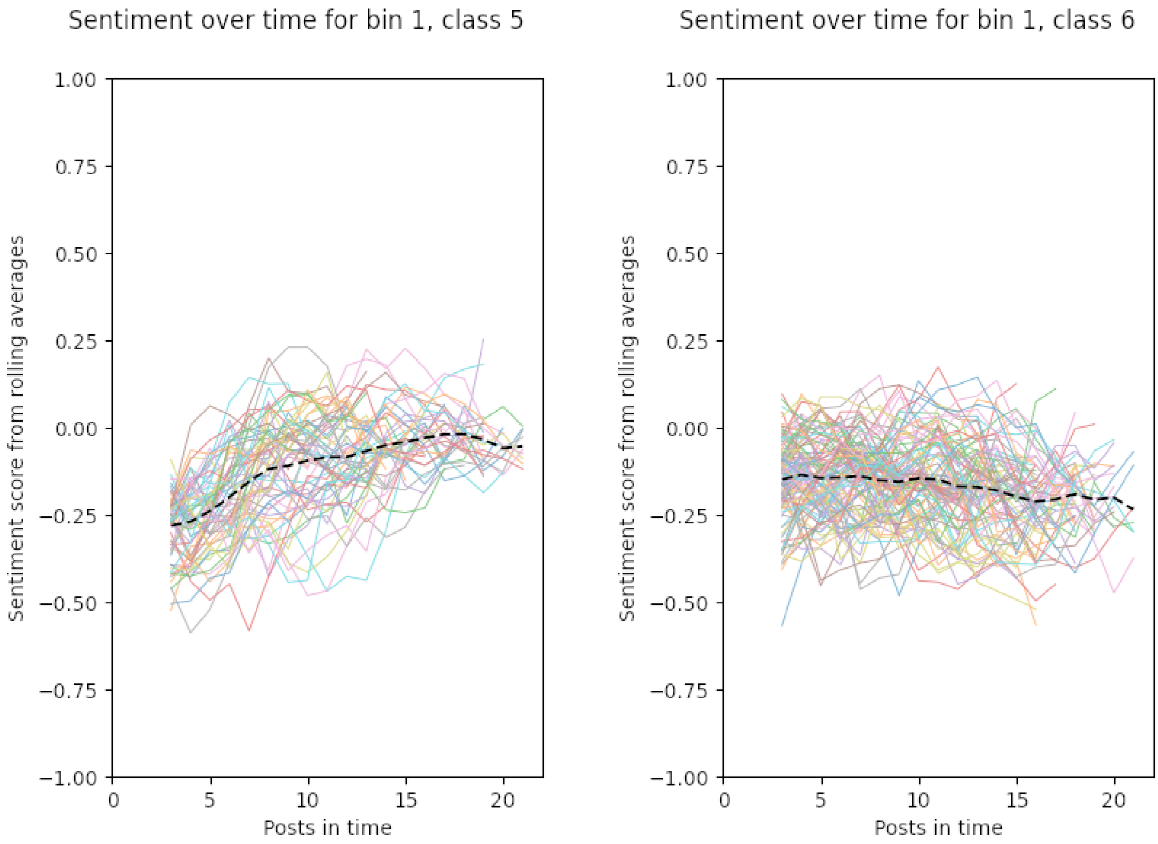


Figure 5.6: Best time series clusters for bin 1 (lengths 7-22), part 2

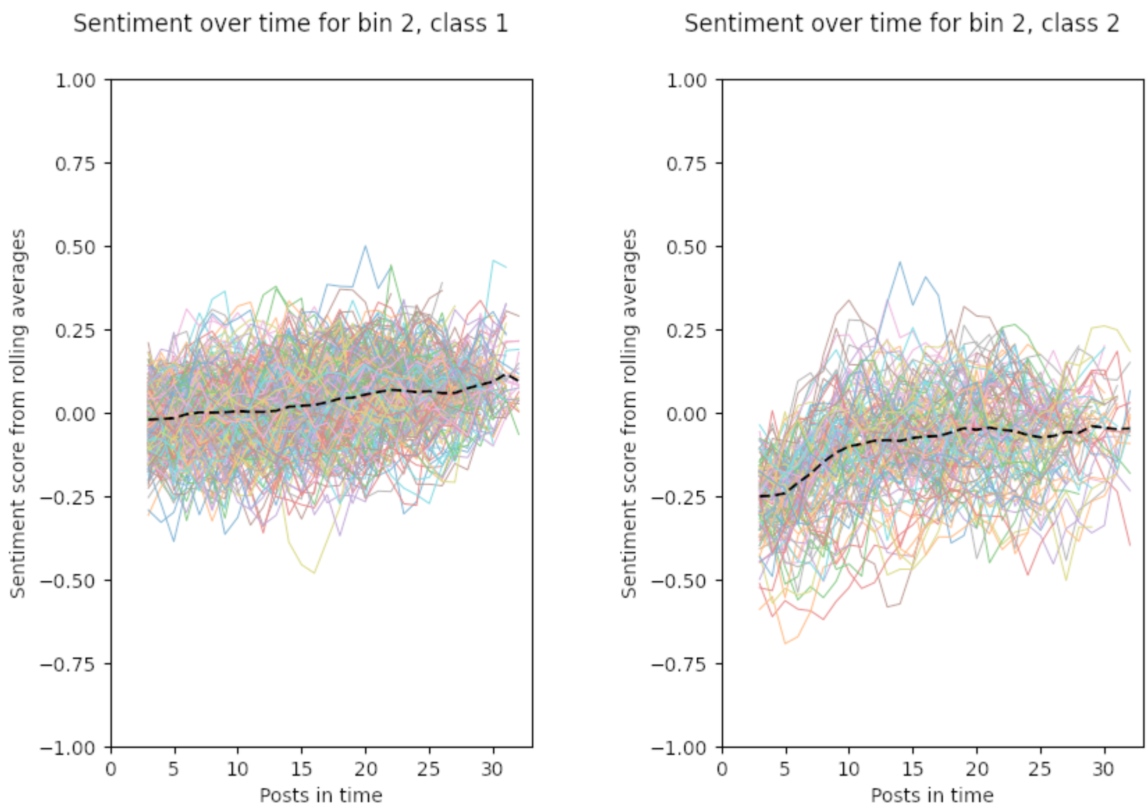


Figure 5.7: Best time series clusters for bin 2 (lengths 23-33), part 1

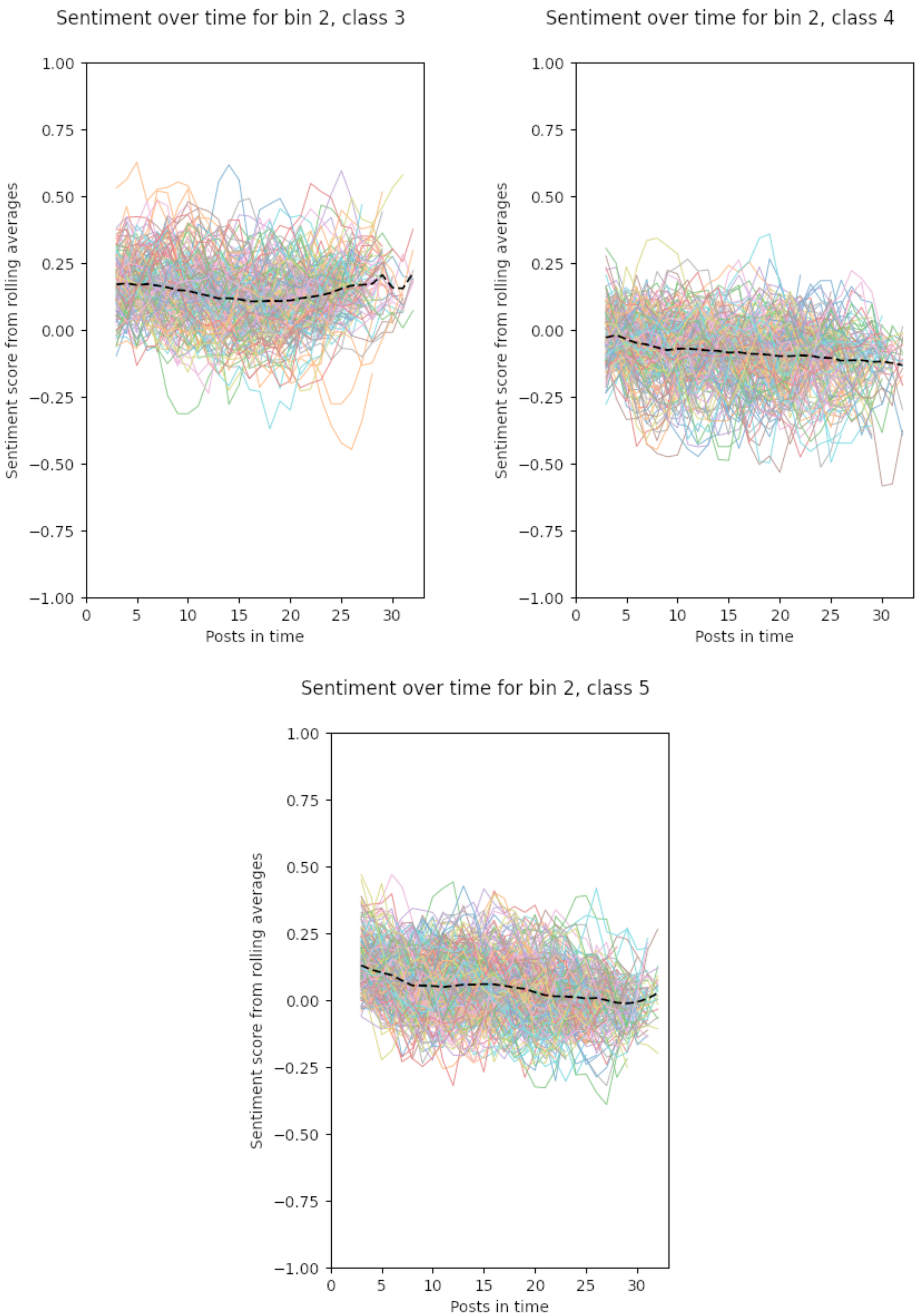


Figure 5.8: Best time series clusters for bin 2 (lengths 23-33), part 2

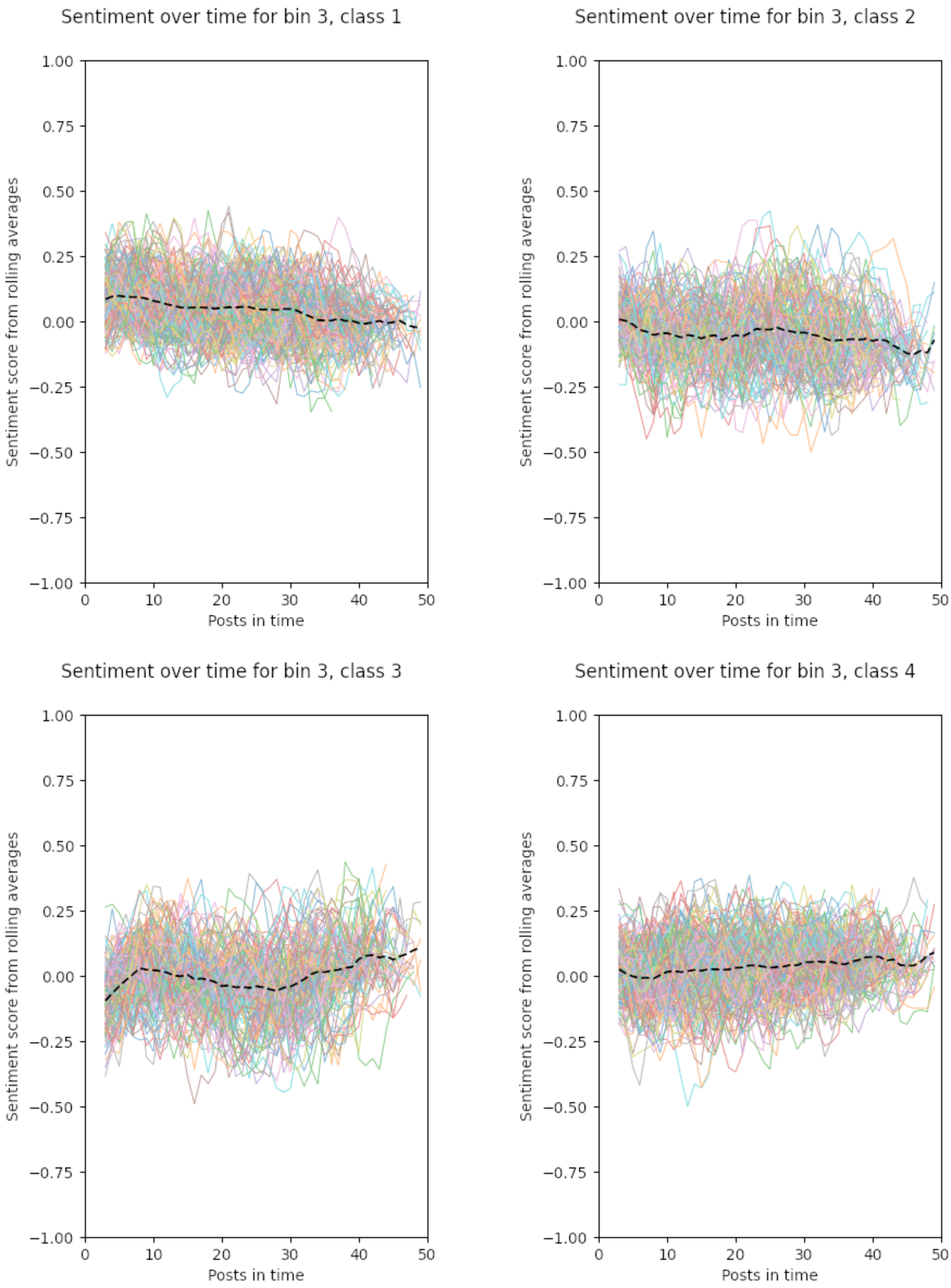


Figure 5.9: Best time series clusters for bin 3 (lengths 34-50), part 1

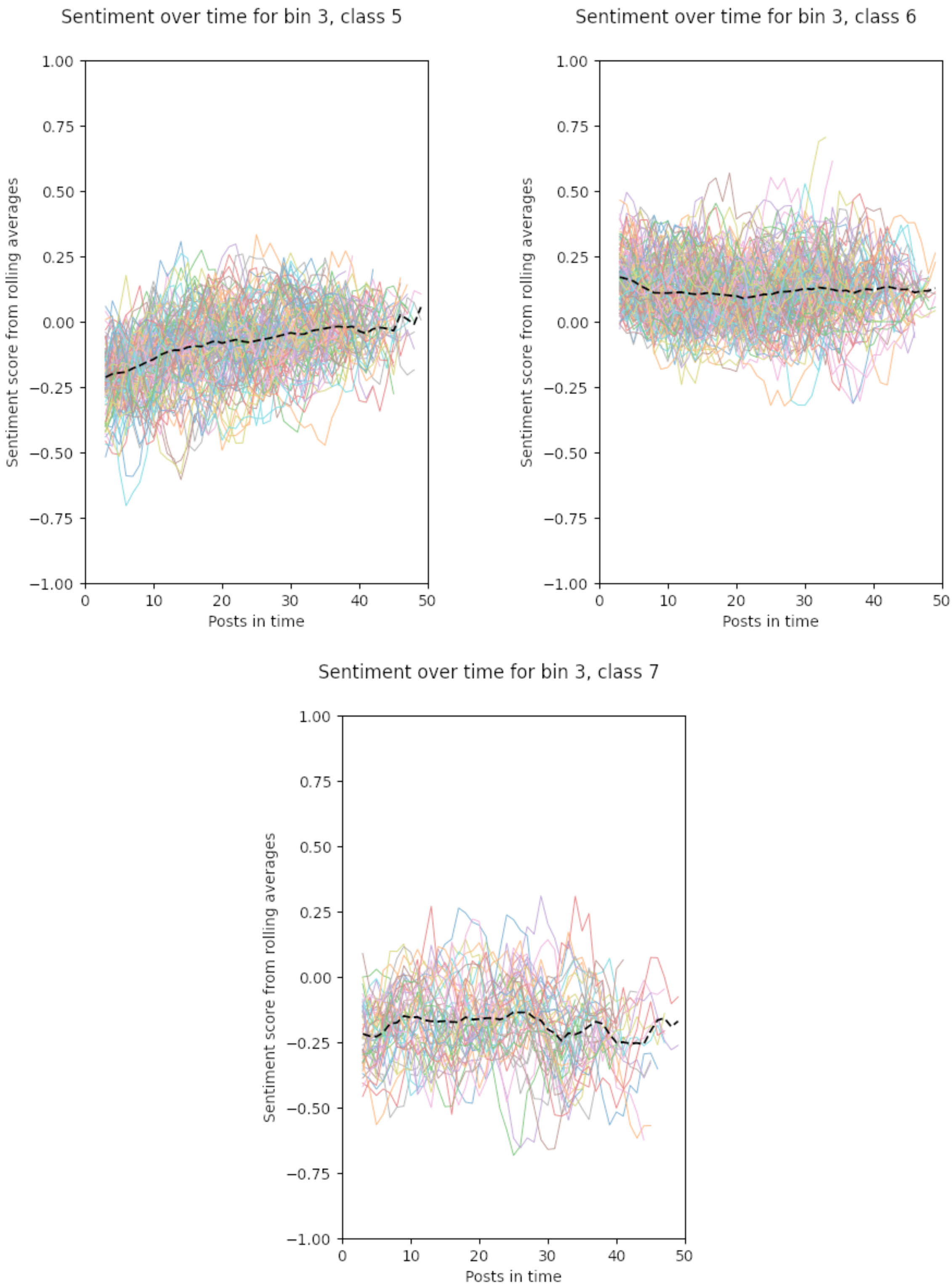


Figure 5.10: Best time series clusters for bin 3 (lengths 34-50), part 2

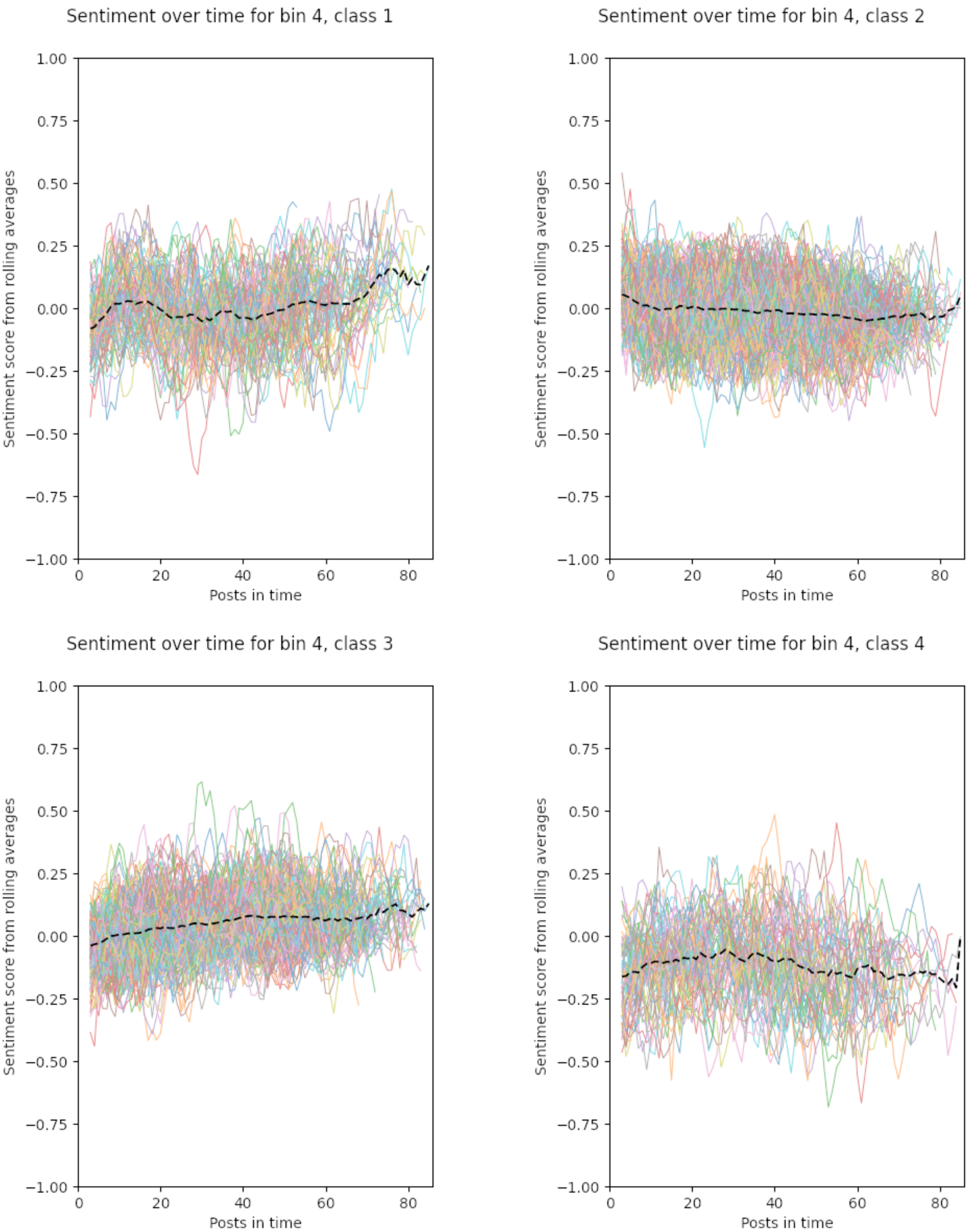


Figure 5.11: Best time series clusters for bin 4 (lengths 51-86), part 1

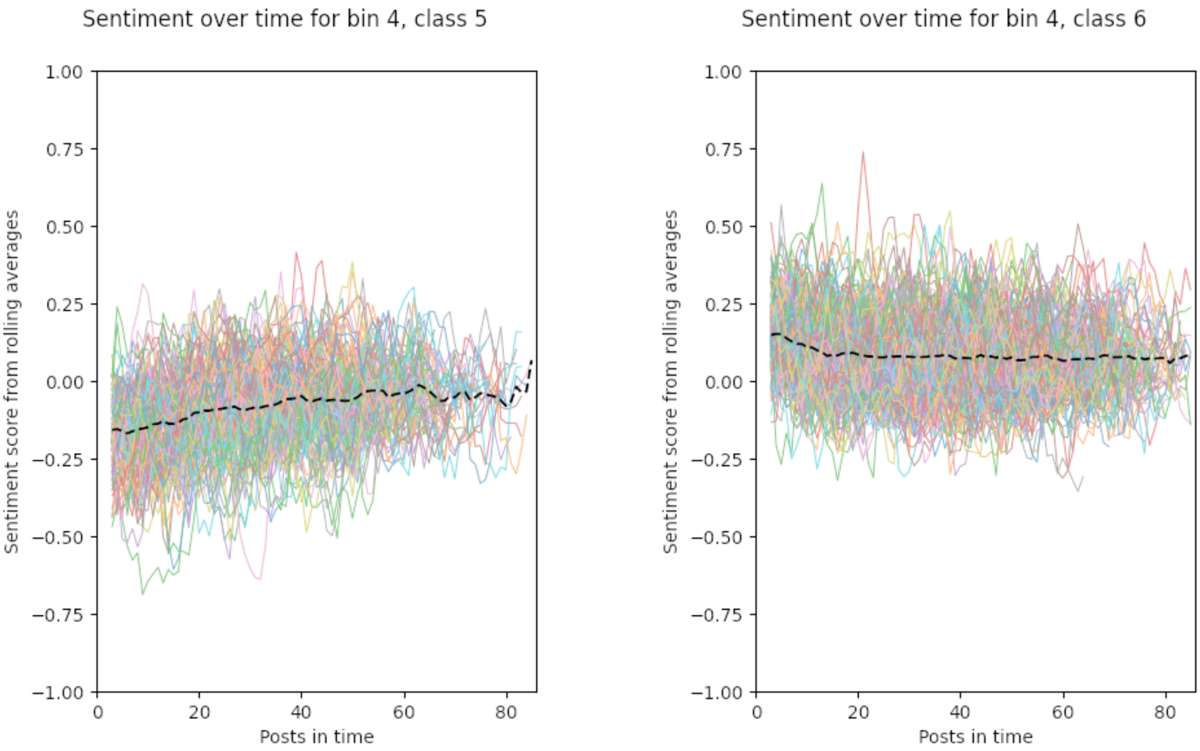


Figure 5.12: Best time series clusters for bin 4 (lengths 51-86), part 2

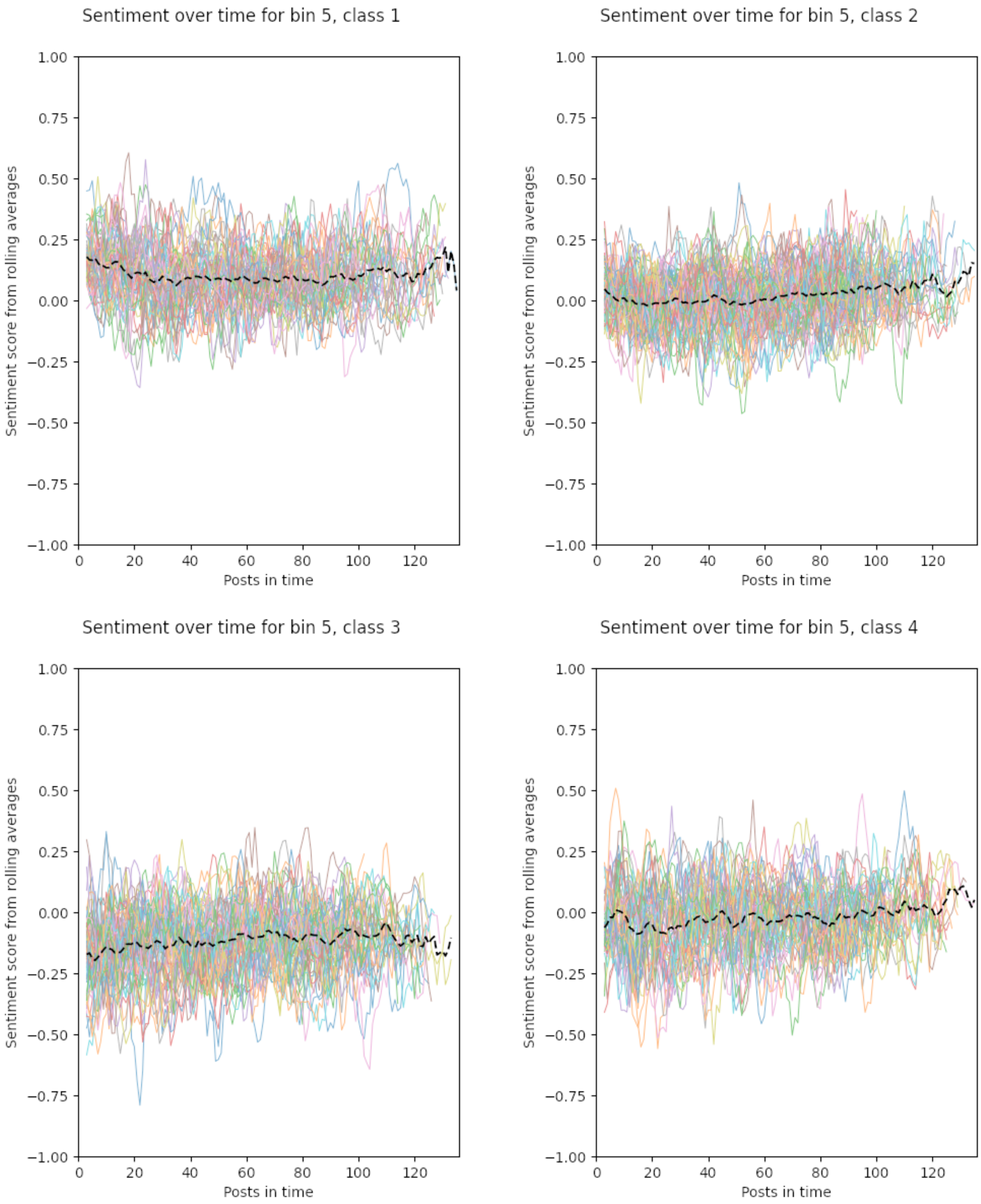


Figure 5.13: Best time series clusters for bin 5 (lengths 87-136), part 1

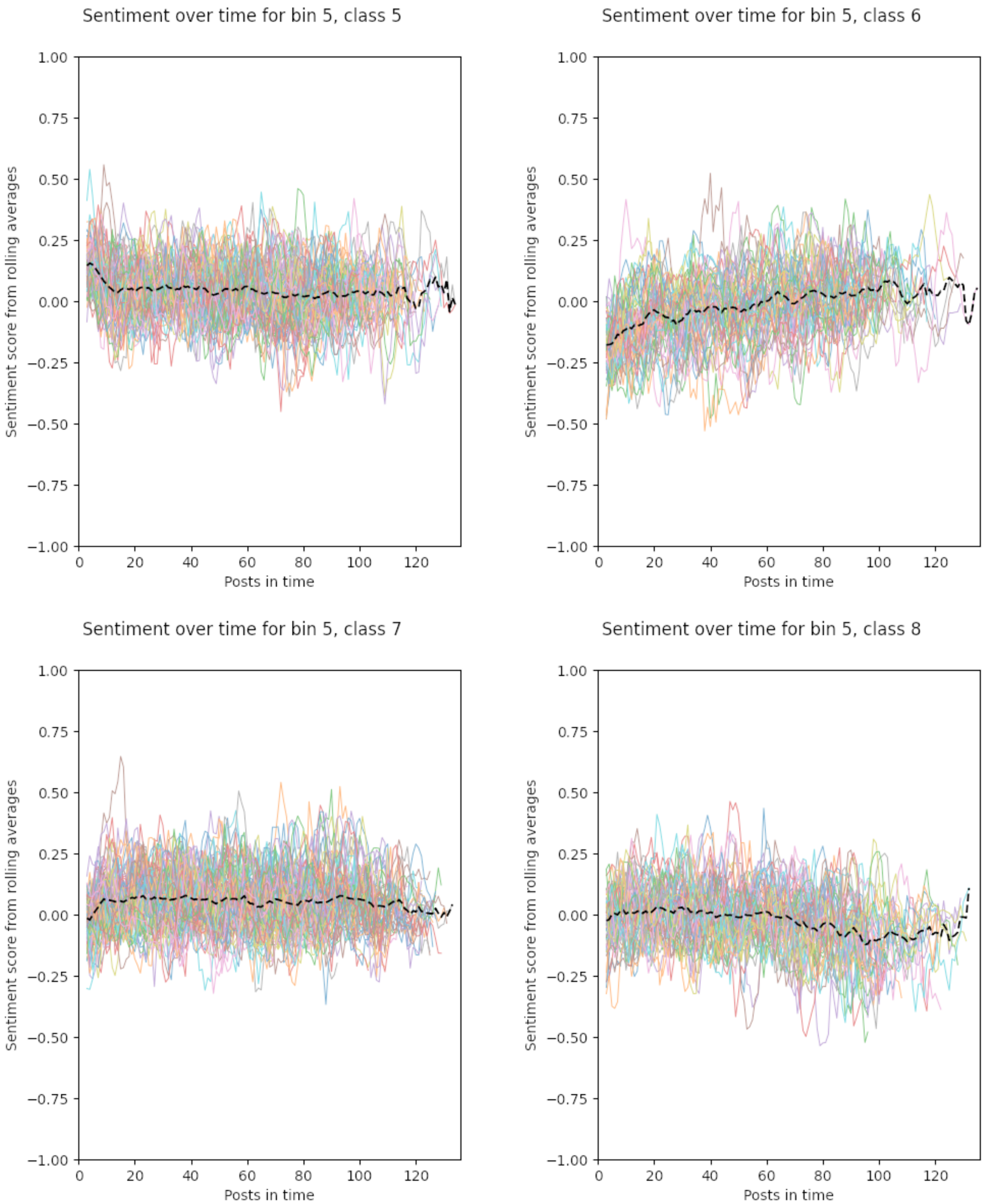


Figure 5.14: Best time series clusters for bin 5 (lengths 87-136), part 2

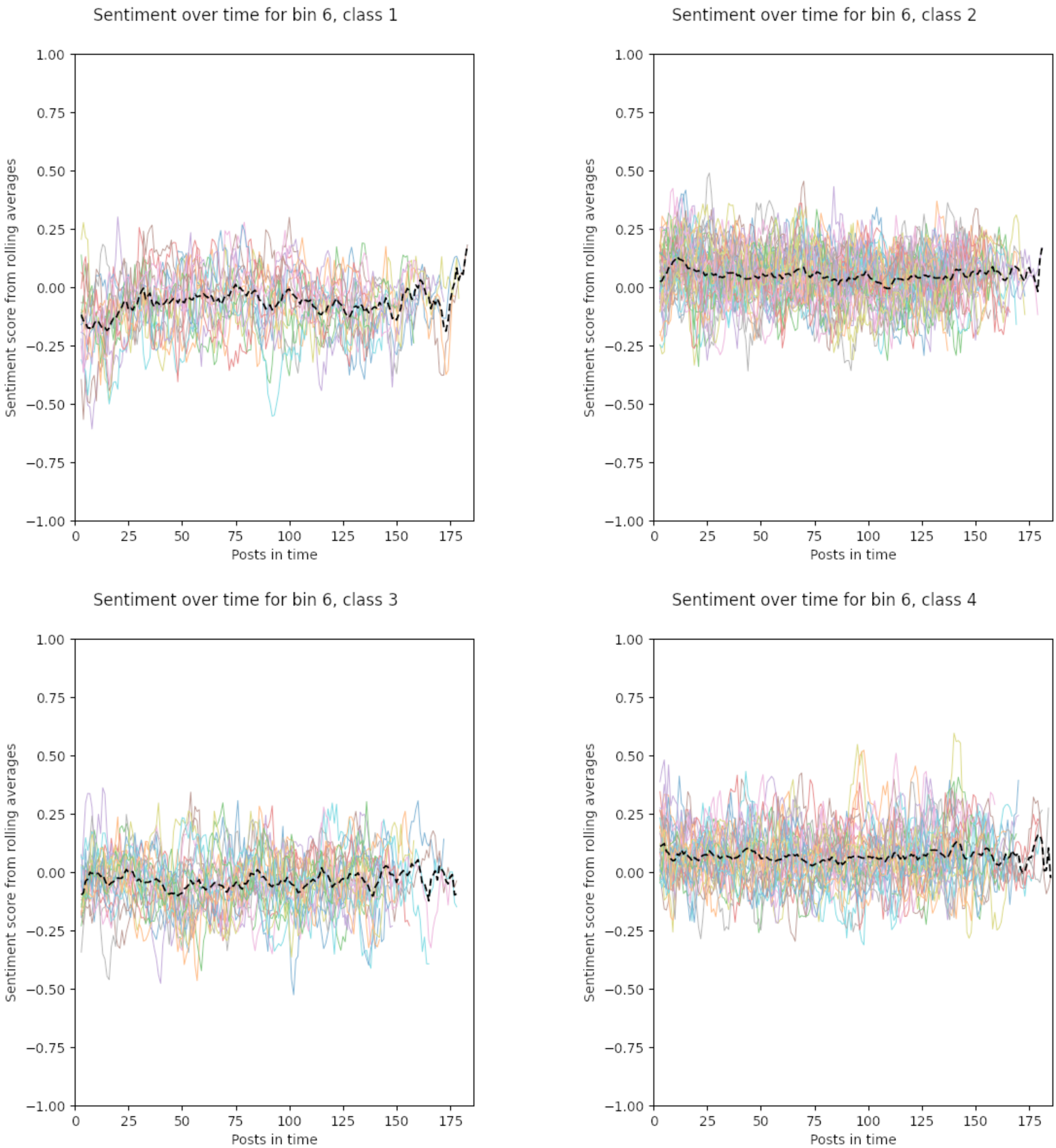


Figure 5.15: Best time series clusters for bin 6 (lengths 137-186), part 1

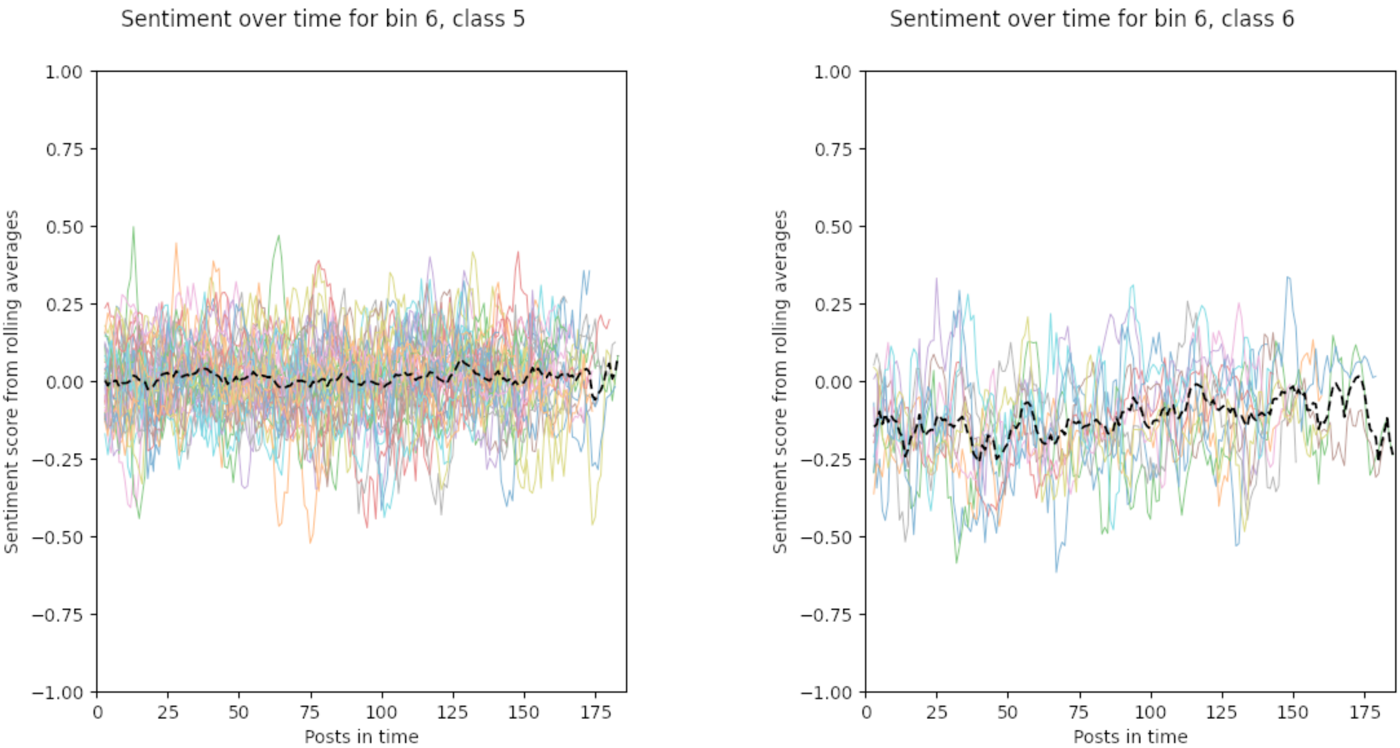


Figure 5.16: Best time series clusters for bin 6 (lengths 137-186), part 2

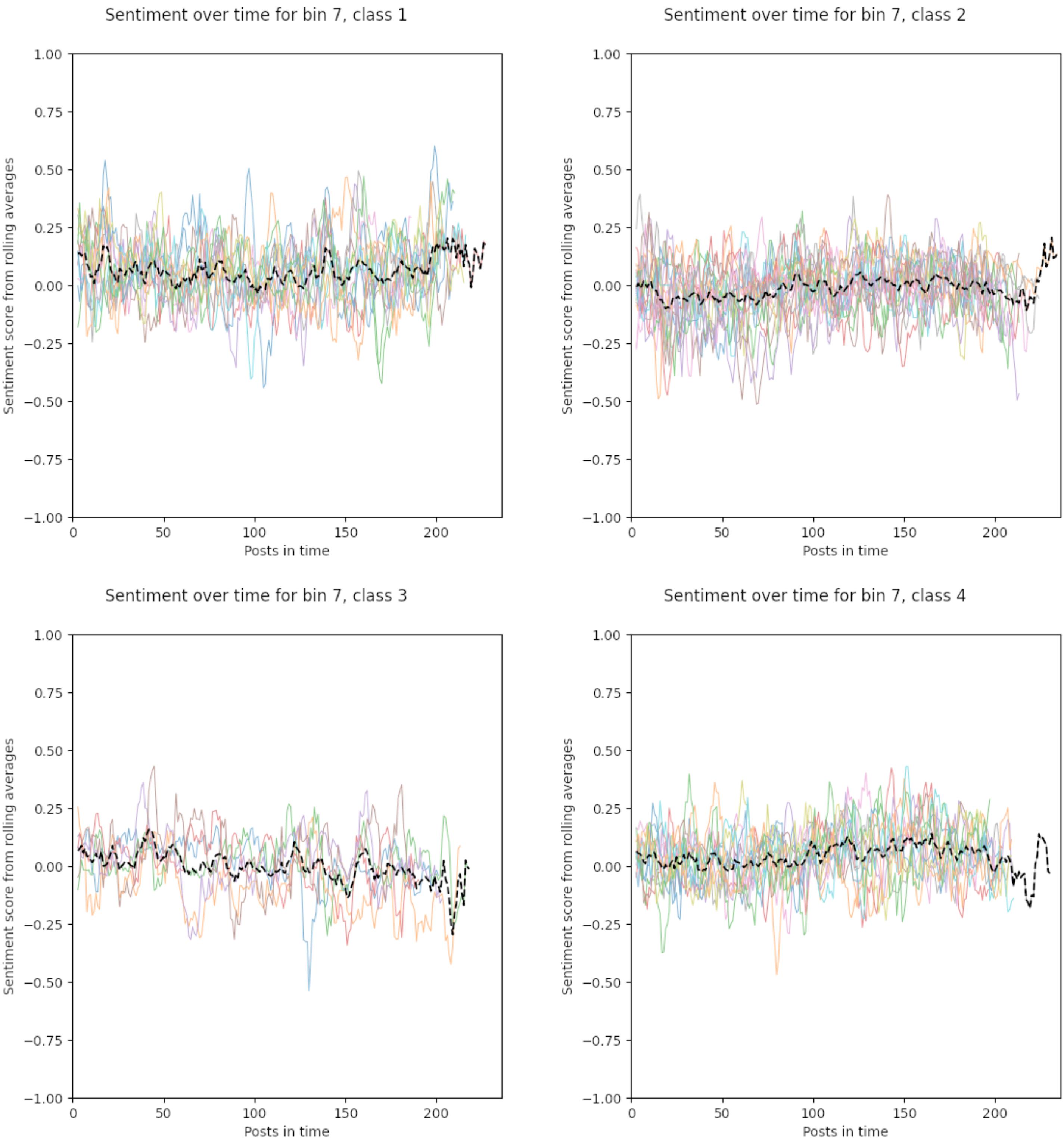


Figure 5.17: Best time series clusters for bin 7 (lengths 187-236), part 1

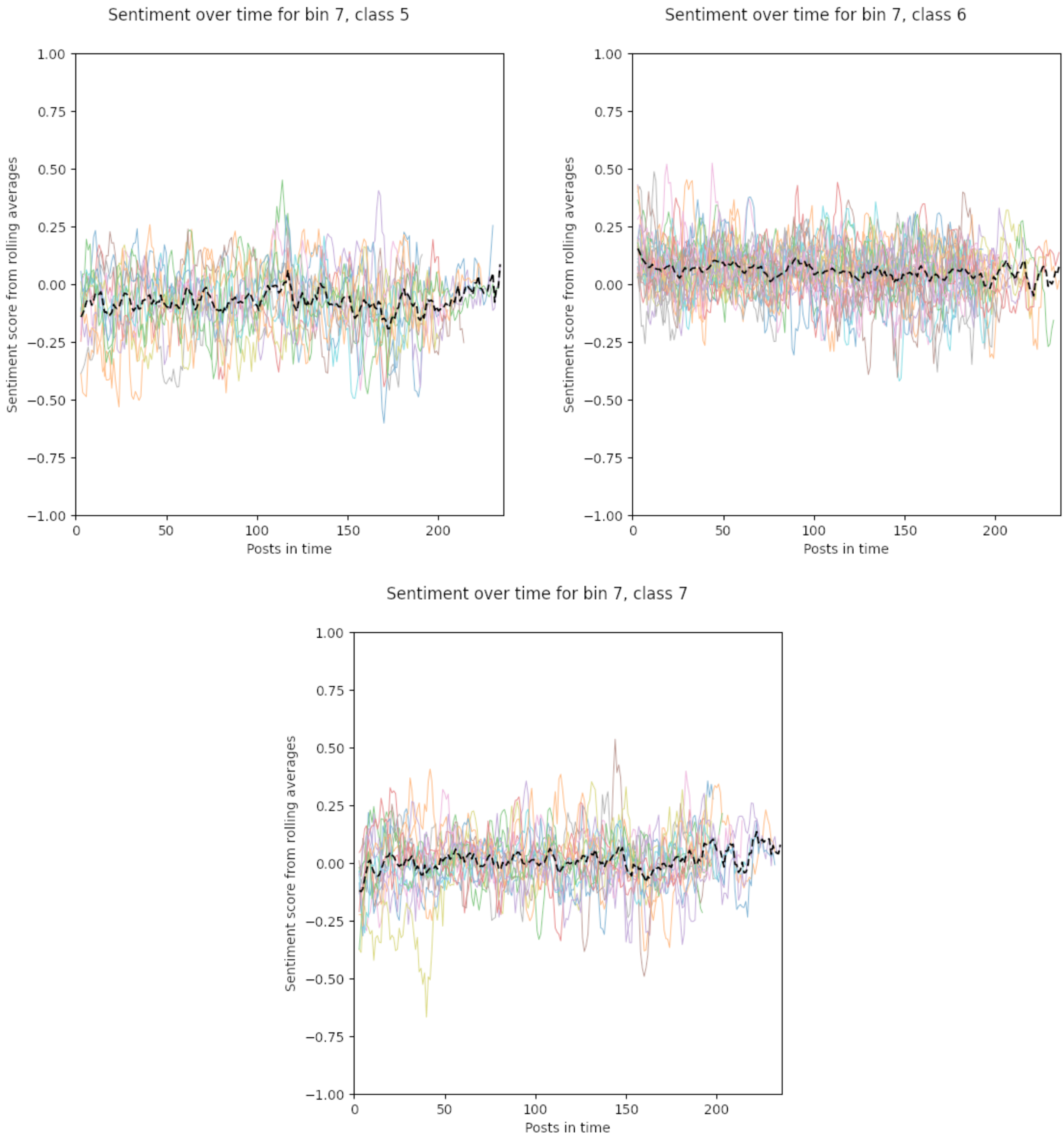


Figure 5.18: Best time series clusters for bin 7 (lengths 187-236), part 2

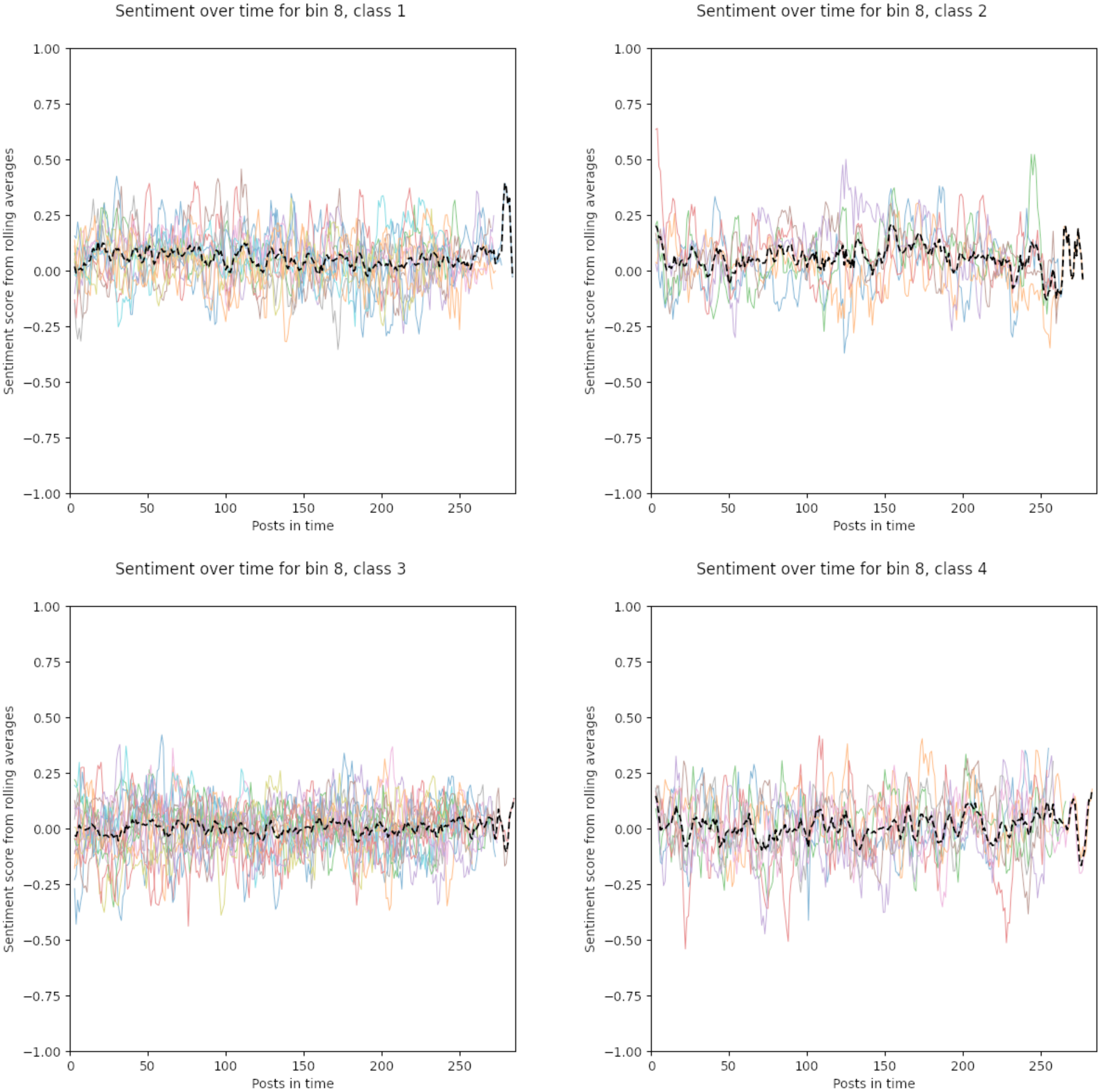
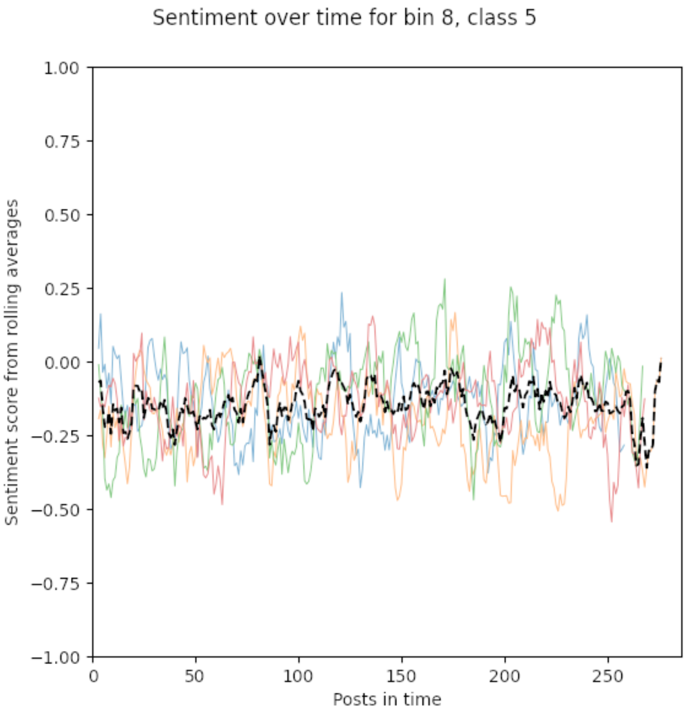


Figure 5.19: Best time series clusters for bin 8 (237-286), part 1



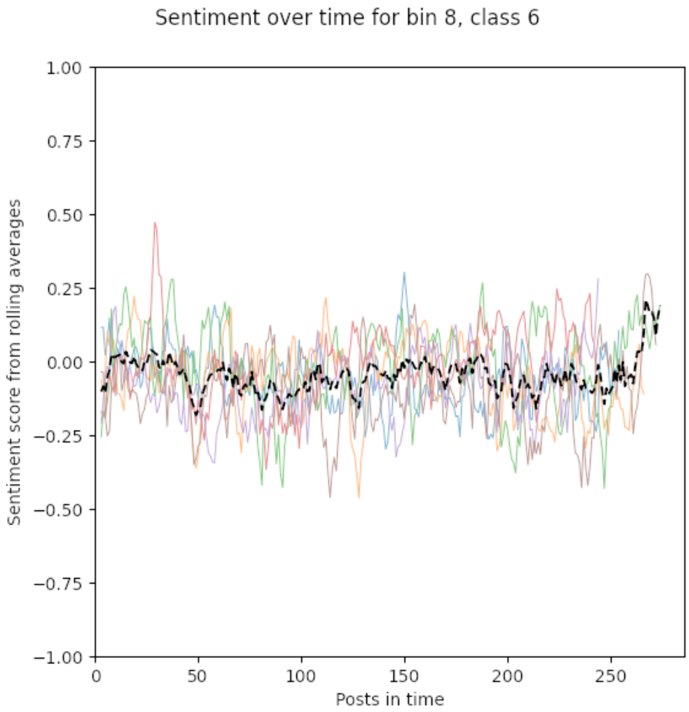


Figure 5.20: Best time series clusters for bin 8 (237-286), part 2

5.5 Discussion

The results imply that there is a lot of variance in the expressed sentiments within discussions. This variance causes the overall sentiment per discussion to average out around a neutral sentiment, without extreme averages, neither positive nor negative. Angry posts are for instance averaged out by more positive posts and enthusiastic posts by negative posts. This is also reflected in the trends of expressed sentiments in bins. One can see changes over time, but they are mostly slight and do not meet the higher thresholds (0.5 and -0.5) for either of the two polarities (positive or negative sentiments).

A general observation for the expressed sentiments over time is that shorter discussions show more interesting patterns: they actually show somewhat distinct classes and changes over time while the classes of longer discussions overlap more and are more flat. In both shorter and longer discussions, however, the higher polarities of at least 0.5 and -0.5 are never reached. This could imply that changes in sentiment are short and on a small scale, or it could mean that in longer discussions, contributing authors are milder.

One observation that could be made from the example discussions with a lower, characteristic, and higher sentiment score is that the example with a lower average sentiment score contains some heated posts and is about a controversial topic, namely abortion. The discussion with the characteristic score is about a less controversial topic (stock market) and contains more factual statements, whereas the discussion with a more positive average sentiment is about music, a topic that some people are enthusiastic about. This might hint that the topic correlates to the expressed sentiments in discussions. Another cause might be that the topic words are included in the sentiment analysis, just as they are in the alignment analysis (discussed in Section 4.5). Some topic words, such as killing (score: -0.6597) and crime (score: -0.5423) in Example 5.4.1 about abortion, carry a strong sentiment. Sentiment scores could thus be more about the topic words than the actual sentiments being expressed. This could be improved in future work by removing the topic words in the sentiment analysis.

Several other things can be noted about this analysis. First of all, the dataset has a drawback that makes sentiment analysis less effective, namely that a textual descriptor has replaced the emojis. Using emojis alongside text could improve the performance of sentiment analysis (Grover, 2022), but the textual descriptors are not recognized by VADER. As the authors of the dataset did not create a conversion list of the emojis to the textual descriptors, we cannot be sure which emoji was used for each descriptor. For future work, the authors could be asked to provide this list or use another dataset that does include emojis, as that could provide clearer or more convincing results.

Furthermore, the metric used for calculating sentiment scores, VADER, is not perfect, and we calculate the scores for posts by averaging the scores of the sentences in the posts. This average might not reflect the overall sentiment that is expressed in a post. In addition, sentiment could change within a post, which is currently averaged out. In future work, one could look at positivity and negativity as separate trends, inspired by what was also done by Wen et al. (2014). Another drawback of VADER is that it does not register sarcasm, which is visible in the last message of Example 5.4.2 and was also discussed by Grover (2022). Different sentiment analysis measures that address these issues could be used in future research.

Another improvement that could be made in future work is to apply spelling corrections or a measure that incorporates spelling errors. Currently, spelling errors cause words not to be recognized by the sentiment analyzer, which makes the proportions of neutral text higher.

Finally, the moving average can be tweaked, the clustering method used could be improved upon, and the notion of time could be investigated as was also discussed for the alignment analysis in Section 4.5.

5.6 Conclusion

The compound sentiment scores of all messages in discussions were computed with VADER and the average sentiment score per discussion was computed. Percentiles were computed and histograms were created for both all messages and the averages of discussions. With these results, we can now answer **Q2.1**. The average sentiment of all messages is 0.01, which is the same as the average sentiment score of all discussions. There is a big range of expressed sentiments in messages and a smaller range of sentiments in the average sentiment scores of discussions.

Discussions were then divided into bins, in the same division as in the alignment analysis. For each discussion, the rolling averages were computed for the sentiment scores. Each bin was clustered to find patterns within the sentiment changing over time for discussions. With that, we can answer **Q2.2**. There are different ways in which sentiment changes over posts in discussions. Different classes were found per bin of discussion lengths, changing from slightly towards one polarity to the other one, but also going back and forth, where the trends mostly lie around a neutral sentiment. The changes in the trends are not large: the thresholds of positive polarity (0.5) and negative polarity (-0.5) are never reached.

Chapter 6

Interplay alignment & sentiment

This chapter discusses the final steps in answering the main research question and investigates the interplay between sentiment, alignment, discussion length, and topic. It answers the following subquestions:

Q3.1 How do the average alignment and the average sentiment of discussions relate?
Q3.2 How does the discussion length relate to the average alignment and the average sentiment in discussions?
Q3.3 How do the alignment and the expressed sentiment trends over time of discussions relate?
Q3.4 How does the topic of a discussion relate to the sentiment and alignment trends over time of discussions?

These questions help answer the main question, with which we aim to understand more of human behavior. If we find relationships, this could have implications for how we design user experiences with for instance conversational agents, and how we implement language models.

6.1 Preprocessing of topic for interplay analysis

For Q3.4, the annotation of the topic per discussion was needed from the dataset. From all original discussions in the dataset, 2894 discussions were annotated by topic, see Section 3.1. However, the general preprocessing already eliminated discussions, and by limiting the discussion length of discussions that were inspected with the bins, not all annotated discussions were included.

A list with the topic per discussion was extracted from the dataset. For each of the discussions included in the alignment and sentiment clustering, the topic was then extracted from this list. This resulted in a list with all included discussions with their annotated topic.

In total, 1602 discussions out of 5136 included discussions were annotated. Table 6.1 shows how many discussions were annotated in each bin of discussion lengths.

Bin	#discussions annotated	#discussions not annotated
1	276	813
2	295	754
3	305	753
4	371	681
5	216	340
6	69	105
7	44	62
8	26	26

Table 6.1: Number of discussions annotated with topic per bin

6.2 Methods

A notebook with the code for the methods that follow can be found at the Github repo[1], see step 4: interplay analysis in the README.

In this thesis, the terms "bin", "alignment class", and "sentiment class" are continuously used. For a reminder of what is meant by these terms, see the following infographic (Figure 6.1):

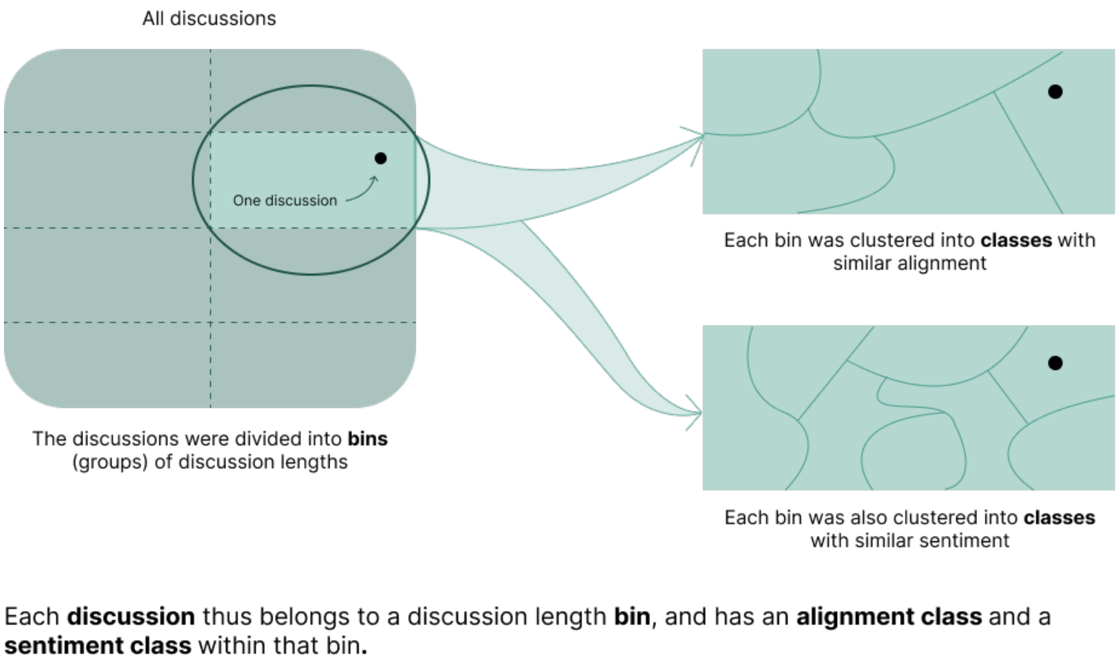


Figure 6.1: What is meant by bins and classes

[1] https://github.com/SuzannaWentzel/Sentiment-Alignment-Interplay

6.2.1 Correlation alignment, sentiment & discussion length (Q3.1 & Q3.2)

To find the relation between the average alignment and sentiment of discussions (and to answer Q3.1), we used the necessary data (average alignment, average sentiment, minimum sentiment, maximum sentiment) that was already computed for Chapter 4 and 5. From these results, the Pearson Correlation Coefficient was used to calculate correlations. The Pearson Correlation Coefficient measures the linear correlation between two variables. It returns a number between -1 and 1, where -1 means a perfect inverse linear correlation, 1 means a perfect linear correlation and 0 implies that there is no linear correlation between the two.

The Pearson correlation coefficient was computed for the following combinations:

- Average alignment & average sentiment
- Average alignment & minimum sentiment
- Average alignment & maximum sentiment

Furthermore, in Section 4.5 we discussed that alignment is most likely related to length, and in Section 5.5 we discussed that sentiment also appears to be related to length, so Q3.2 was added, and we investigated more combinations with the Pearson Correlation coefficient:

- Average alignment & discussion length
- Average sentiment & discussion length
- Minimum sentiment & discussion length
- Maximum sentiment & discussion length

Each combination of variables was also plotted in a heat plot showing the joint distribution to visualize potential correlations.

6.2.2 Interplay discussion topic and alignment & sentiment over time (Q3.3 & Q3.4)

From the alignment analysis in Chapter 4, the alignment class per discussion was extracted, which included a mean alignment trend for that class. Similarly, from the sentiment analysis in Chapter 5, the sentiment class per discussion was extracted, including a mean sentiment trend for that class. These were needed to investigate the interplay for Q3.3.

Q3.4 followed from the discussions in Sections 4.5 and 5.5. From the additional preprocessing which was described in Section 6.1, the topic of each discussion was obtained, such that the interplay with the topic could also be investigated.

With this data, contingency tables were created. A contingency table is a cross-tabulation that displays the occurrence count of each combination of two categorical variables. For each discussion length bin, three contingency tables were created:

- Alignment class vs. sentiment class
- Alignment class vs. topic
- Sentiment class vs. topic

Not all discussions in the dataset were annotated with their topic, so for the contingency tables that concerned the topic, only discussions that included the topic were included. This means that there is less data in the contingency tables that concern the topic. The last two bins (bin 7 with 187-236 posts and bin 8 with 237-286 posts) had less than 50 annotations. We have computed these contingency tables and Cramér's V values for completeness nevertheless, but we did not investigate these further.

Some topics in the contingency tables proved to be sparsely populated with discussions, which means that they were not representative and thus not insightful for drawing conclusions. Therefore, topics that included fewer discussions than the number of classes for each bin were removed. This sometimes resulted in classes without any annotated discussions, these classes were also removed for brevity. This resulted in reduced contingency tables.

For each sentiment class (in all bins), the affective shift was computed by subtracting the final sentiment score from the starting sentiment score of the mean trend. Similarly, for each alignment class (in all bins), the alignment change was computed by subtracting the final time-based overlap from the starting time-based overlap of the mean trend. With the affective shift and alignment change, the rows and columns in the contingency tables were ordered.

Cramér's V was computed for each contingency table. Cramér's V can be used to compute the association between two categorical variables. It returns a value between 0 and 1, where 0 means no association and 1 means a complete association (when each variable determines the other).

Combining insights from the contingency tables with the values from Cramér's V, the patterns within and across bins between alignment, sentiment, and topic were investigated.

6.3 Results

6.3.1 Correlation alignment, sentiment & discussion length

The Pearson correlation coefficients between the variables are shown in Table 6.2. The respective joint distributions are plotted in Figures 6.2 - 6.8. The correlation between average alignment and average sentiment is near 0, meaning that there is no linear correlation. The plot (see Figure 6.4) looks like a joint normal distribution. For the average alignment and the minimum sentiment and maximum sentiment, the correlations are respectively -0.23 and 0.27, meaning low linear correlations. The plotted joint distributions (Figures 6.2 & 6.3) show that as alignment gets higher, the minimum sentiment gets lower and the maximum sentiment gets higher.

The correlation between alignment and discussion length is 0.7, which indicates somewhat of a linear correlation. This is confirmed in the joint distribution in Figure 6.8, though it shows to not be linear but curved, where longer discussions have a higher average overlap.

The correlation for the average sentiment and discussion length is -0.03, which means that there is no linear correlation. Looking at the joint distribution in Figure 6.8, one sees that the average sentiment of longer discussions is more around 0 (neutral sentiment). The correlations for discussion length and minimum sentiment and maximum sentiment are respectively -0.39 and 0.39, which means low linear correlation. The distributions (see Figures 6.6 & 6.7) show that as discussions get longer, the minimum and maximum sentiments get more extreme.

		Pearson Correlation Coefficient
Average alignment	*Average sentiment*	0.0012
Average alignment	*Minimum sentiment*	-0.2277
Average alignment	*Maximum sentiment*	0.2678
Average alignment	*Discussion length*	0.7195
Average sentiment	*Discussion length*	-0.0285
Minimum sentiment	*Discussion length*	-0.3854
Maximum sentiment	*Discussion length*	0.3937

Table 6.2: Pearson correlation per combination of variables

6.3.2 Interplay sentiment over time & alignment over time

The Cramér's V values of sentiment against alignment for each bin of discussion length are shown in Table 6.3. The accompanying contingency tables can be found in Appendix F.1. The Cramér's V values for bins 1 - 4 (all near 0) show that there is no association at all between alignment and sentiment classes.

For the higher bins, there is an increasing association, though they are also sparser so this pattern might be a result of overfitting.

	Bin 1	*Bin 2*	*Bin 3*	*Bin 4*	*Bin 5*	*Bin 6*	*Bin 7*	*Bin 8*
Cramér's V	0.0762	0.0899	0.0934	0.0677	0.1448	0.1992	0.2646	0.3287

Table 6.3: Cramér's V for the sentiment and alignment classes

The contingency tables show that they are almost randomly uniformly distributed. Some classes are less populated than others, but overall there is no clear division. Locally, however, some interesting things can be seen. One example in bin 1 (Table F.1) is that class 2 of alignment trends and class 1 of sentiment trends appear to correspond. These trends are shown in Figure 6.9, which shows a flatter class of alignment, and a flatter, slightly decreasing class of sentiment.

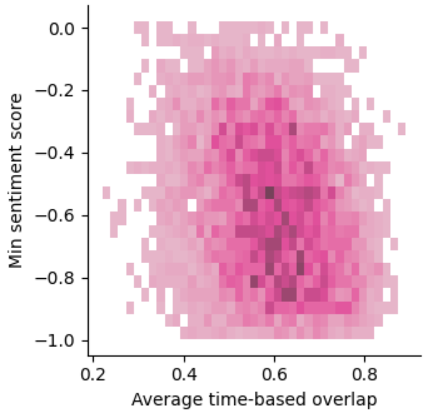


Figure 6.2: Joint distribution of average alignment & min sentiment

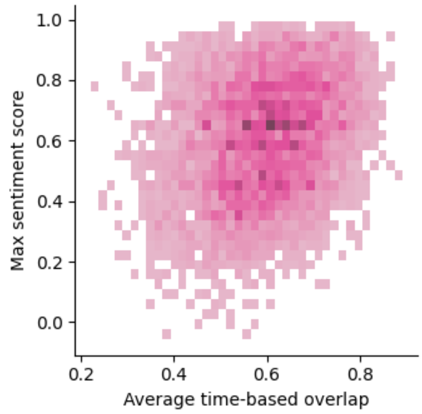


Figure 6.3: Joint distribution of average alignment & max sentiment

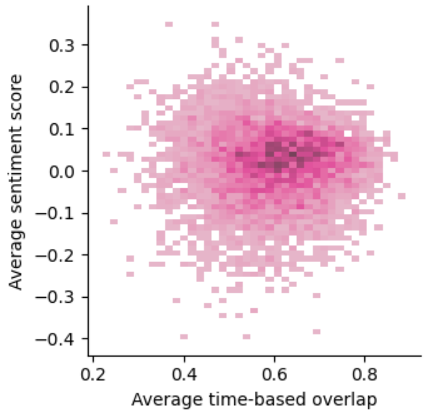


Figure 6.4: Joint distribution of average alignment & average sentiment

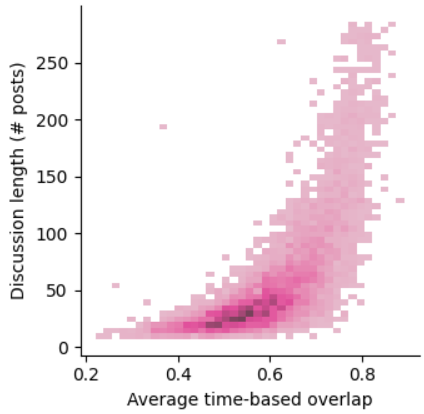


Figure 6.5: Joint distribution of average alignment & discussion length

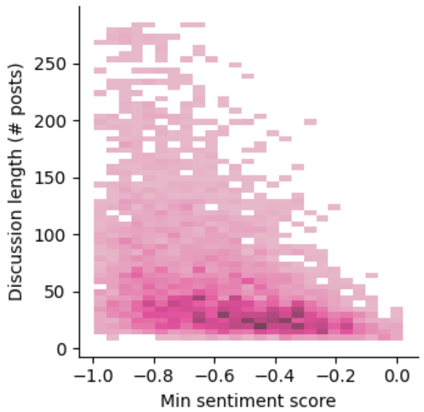


Figure 6.6: Joint distribution of minimum sentiment & discussion length

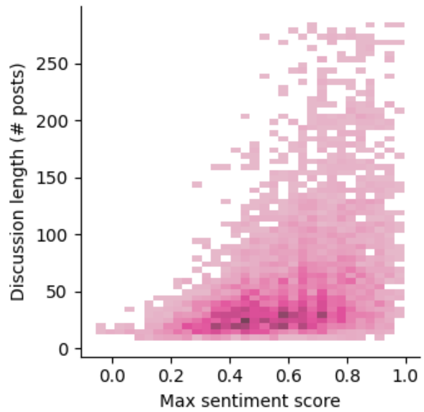


Figure 6.7: Joint distribution of maximum sentiment & discussion length

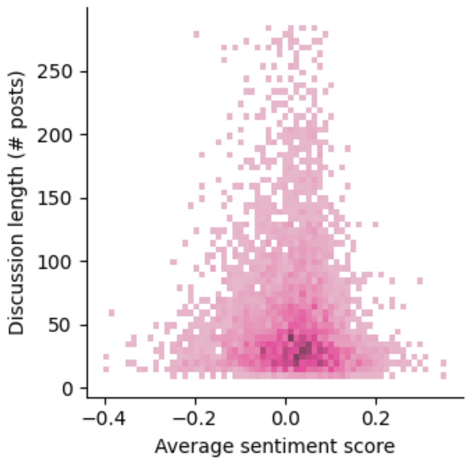


Figure 6.8: Joint distribution of average sentiment & discussion length

6.3.3 Interplay alignment over time & topic

The Cramér's V values of each bin of discussion lengths can be found in Table 6.4. It shows both the Cramér's V values with and without the sparse data removed. The Cramér's V values show more association between the alignment classes and the topics than between the alignment classes and the sentiment classes, though the associations are not high.

	Bin 1	*Bin 2*	*Bin 3*	*Bin 4*	*Bin 5*	*Bin 6*
Cramér's V for all data	0.2067	0.2283	0.2081	0.1795	0.2827	0.3437
Cramér's V without sparse data	0.1889	0.1774	0.1955	0.1413	0.1849	0.3219

Table 6.4: Cramér's V for the alignment classes and topics

The ordered contingency tables without the sparse data (per bin of discussion length) are shown in Tables 6.5 - 6.10. The contingency tables of the last two bins can be found in Appendix F.2. The original contingency tables where sparse data is not removed can be found in Appendix F.3.

The contingency tables should be read as follows. Each cell shows how many discussions have that combination of alignment class (group of alignment trends) and topic. The rows are ordered ascending on the alignment change. A more saturated cell shows that it is highly populated.

The contingency tables show that some topics are highly represented by an alignment class, such as "evolution" by class 2 in bin 1 (Table 6.5), in bin 2 (Table 6.6) "abortion" by class 2, and "evolution" by class 1 in bin 3 (Table 6.7). Especially this latter example stands out from the contingency table. To illustrate this example, the alignment class is plotted in Figure 6.10, a class with a steeper alignment increase.

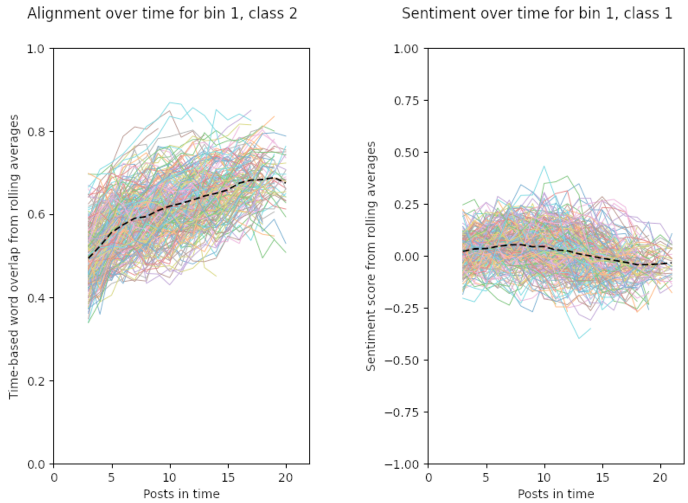


Figure 6.9: Example of an alignment class and a sentiment class that often go together in bin 1

An excerpt of a discussion from this example class with the topic "evolution" is given in Example 6.3.1. Here one can see a steep increase in alignment starting lower and increasing rapidly. Note that Figure 6.10 shows the time-based overlap from the moving averages, whereas the excerpt gives the raw time-based overlap scores.

Example 6.3.1: Discussion about "evolution" which is highly represented by the class in Figure 6.10

...

Message 1 (author 1164): I was reading my bible and decided to take a spin through creation. For a long time since I was a kid I noticed genesis was out of order. Not even according to science, but to its self. There seems to be more than one description of creation. Which makes you wonder did mosses understand what God was imparting. (...)

Message 2 (author 60): Thefirst 5 books of the bible contain two texts in it, the J and E texts which represent two different traditions and legends of two different peoples which ended up becoming combined to form the Jews. However as these are both the word of god they were both included when it was writen down for fearing of modifing the Word of God. (...) (Time-based overlap: **0.28**)

Message 3 (author 401): No no no........ Pre-Creation week: The earth is without form, and consists of water alone. Day 1: God creates light. He separates the light from the dark and names the light day, and the dark night. Day 2: God creates the sky separating the unformed Earth with the rest of the unformed universe. (...) (Time-based overlap: **0.36**)

...

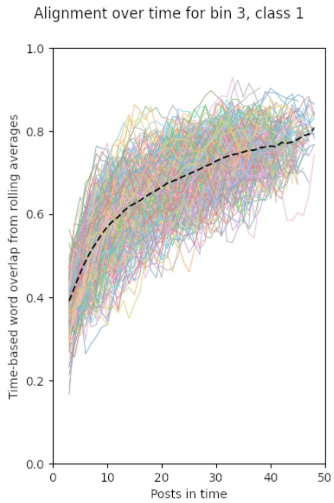


Figure 6.10: The topic "evolution" was highly represented by this alignment class

Message 12 (author): how is either Genesis 1 or 2 in error? it looks to me as though Gen 1 is an overview of the creation week, and Gen 2 brings the specific creation of man and woman into focus. how is that error? maybe I dont see it... (Time-based overlap: **0.6**)
Message 13 (author): Because both have different creation orders. Even if chapter 2 is 'specific' that doesn't even address the problem that the OP presented. Basically, what is the order of creation-chapter 1 or 2? Please justify your answer. (Time-based overlap: **0.57**)

Some topics are more spread across the alignment classes, such as "abortion" and "gun control" in bin 1 (Table 6.5), "evolution" and "gay marriage" in bin 2 (Table 6.6), and "existence of God" in bin 3 (Table 6.7). It changes per bin which topics are more spread, no topic shows such a pattern across the bins of discussion lengths.

The contingency tables show that some topics belong to multiple classes with different alignment patterns, such as "abortion" and "gun control" in bin 2 (Table 6.6), "abortion" in bin 3 (Table 6.7), and "abortion", "evolution" and "gun control" in bin 4 (Table 6.8). To illustrate, the different alignment patterns relating to "abortion" in bin 4 (class 2 and 5) are shown in Figure 6.11. As can be seen, class 5 is much steeper than class 2.

Inspecting the results the other way around, we see that few classes are highly represented by one topic. Some examples are "gun control" for classes 5 and 2 in bin 3 (Table 6.7) and "gun control" for alignment class 4 in bin 6 (Table 6.10).

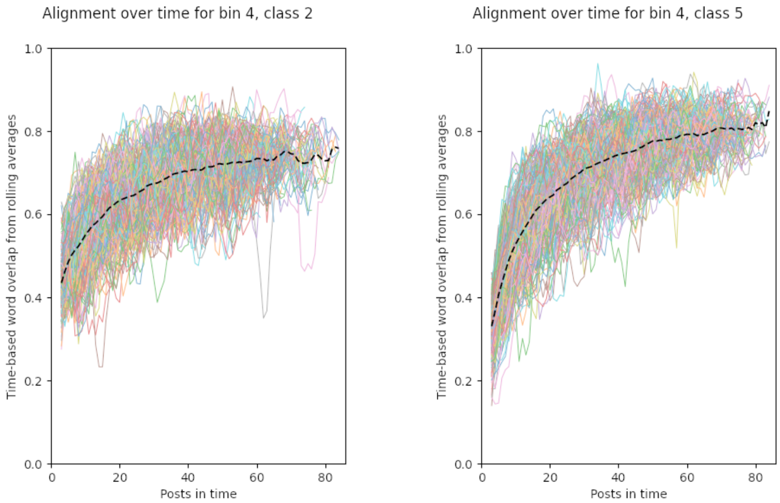


Figure 6.11: The topic "abortion" was highly represented by these alignment classes

Looking at patterns across bins, one can see that the classes that fit "gun control" best often start with the lowest alignment. Another pattern is that the classes most often seen in discussions about "abortion" mostly start with an alignment around 0.4 and reach 0.6 around 20 posts, indicating not a steep nor a flat increase of alignment but something in between. Some other topics show patterns across a few bins, but not the rest. An example is "gay marriage", which corresponds in some bins to classes with high alignment changes, but in the other bins, their classes are more spread out.

		Topics							**Alignment changes**
		abortion	*climate change*	*communism vs capitalism*	*evolution*	*existence of God*	*gay marriage*	*gun control*	
Alignment classes	*4*	7	3	0	19	2	4	21	0.1454525
	2	13	1	2	35	1	8	17	0.1810517
	1	12	0	2	14	1	2	30	0.2037299
	3	14	4	0	19	3	9	29	0.2978775

Table 6.5: Contingency table showing the interrelation between alignment classes and topics, bin 1 (7-22 posts)

		Topics						Alignment change
		abortion	climate change	evolution	existence of God	gay marriage	gun control	
Alignment classes	2	17	1	16	0	2	19	0.18047
	4	5	3	9	0	3	16	0.216366
	5	5	1	20	4	7	7	0.222946
	6	10	5	14	3	11	30	0.329917
	1	10	0	17	3	10	14	0.370287
	3	2	1	4	1	3	10	0.384073

Table 6.6: Contingency table showing the interrelation between alignment classes and topics, bin 2 (23-33 posts)

		Topics								Alignment changes
		abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gun control	
Alignment classes	2	9	3	0	2	12	3	5	21	0.2201436
	4	15	0	1	0	16	6	4	10	0.2265328
	5	4	0	1	0	5	0	2	14	0.326283
	1	16	2	1	1	42	5	10	18	0.4169202
	3	12	3	5	2	18	3	8	20	0.4294556

Table 6.7: Contingency table showing the interrelation between alignment classes and topics, bin 3 (34-50 posts)

		Topics						Alignment change
		abortion	communism vs capitalism	evolution	existence of God	gay marriage	gun control	
Alignment classes	2	19	0	20	7	10	25	0.3238632
	1	15	1	33	6	13	13	0.339139
	4	7	0	3	0	1	5	0.3474578
	3	15	2	19	2	7	32	0.4353458
	5	26	2	27	4	10	32	0.5182979

Table 6.8: Contingency table showing the interrelation between alignment classes and topics, bin 4 (51-86 posts)

		Topics						Alignment change
		abortion	death penalty	evolution	existence of God	gay marriage	gun control	
Alignment classes	4	13	2	7	3	5	11	0.258296
	3	5	2	3	0	3	6	0.417103
	2	24	2	28	10	6	15	0.428433
	1	18	1	15	0	7	17	0.548944

Table 6.9: Contingency table showing the interrelation between alignment classes and topics, bin 5 (87-136 posts)

		Topics				Alignment change
		abortion	*evolution*	*gay marriage*	*gun control*	
Alignment classes	*3*	1	0	0	1	0.282995
	5	5	1	0	1	0.315733
	4	4	4	1	10	0.473639
	2	7	10	6	4	0.505759
	1	3	1	2	1	0.600344

Table 6.10: Contingency table showing the interrelation between alignment classes and topics, bin 6 (137-186 posts)

6.3.4 Interplay sentiment over time & topic

The Cramér's V values of each bin of discussion lengths can be found in Table 6.11. It shows both the Cramér's V values with and without sparse data removed. The Cramér's V values show more association between the sentiment classes and the topics than between the sentiment and alignment classes and between the alignment classes and topics. The values show some association, which increases as the discussion lengths increase.

	Bin 1	*Bin 2*	*Bin 3*	*Bin 4*	*Bin 5*	*Bin 6*
Cramér's V for all data	0.3153	0.3350	0.3279	0.3651	0.4072	0.4576
Cramér's V without sparse data	0.2999	0.3106	0.3080	0.3558	0.4478	0.5344

Table 6.11: Cramér's V for the sentiment classes and topics

The ordered contingency tables without sparse data (per bin of discussion length) are shown in Tables 6.12 - 6.17. The contingency tables of the last two bins can be found in Appendix F.4. The original contingency tables (without sparse data removed) are shown in Appendix F.5.

The contingency tables should be read as follows. Each cell shows how many discussions have that combination of sentiment class (group of sentiment trends) and topic. The rows are ordered ascending on their affective shift. A more saturated cell shows a highly populated cell.

The contingency tables show that topics are often highly represented by sentiment classes. In each discussion length bin, some topics and classes stand out. To name some: in bin 2 (Table 6.13) "gay marriage" is for instance mostly represented by sentiment class 1, in bin 4 (Table 6.15), "abortion" is represented by sentiment class 2, and "existence of God" by class 6. To illustrate this last example, sentiment class 6 is shown in Figure 6.12, which shows a decline in sentiment from positive to neutral.

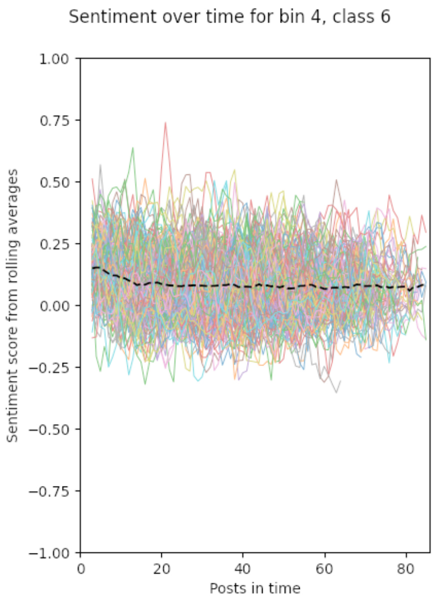


Figure 6.12: The topic "existence of God" was highly represented by this sentiment pattern

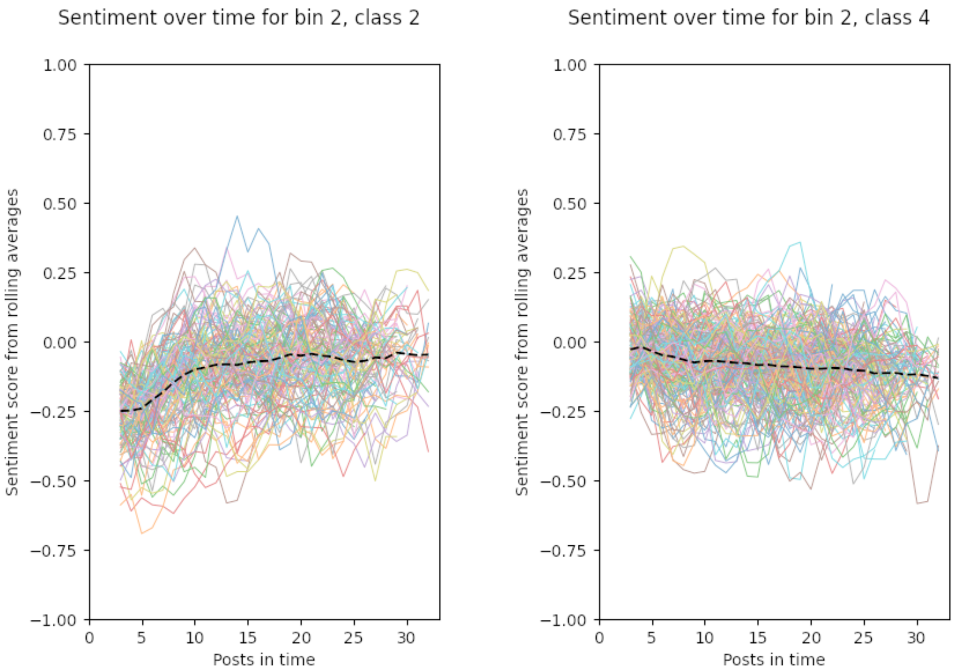


Figure 6.13: Opposing sentiment classes that highly represent the topic "gun control" in bin 2

Some topics are more spread across classes, such as "abortion" in bin 1 (Table 6.12), "climate change" and "existence of God" in bin 2 (Table 6.13), and "gun control" in bin 3 (Table 6.14). Especially "evolution" does not seem to be characteristic of one sentiment class (see bin 1, bin 3, bin 4 (Table 6.15), and bin 5 (Table 6.16)).

Some topics show to correspond to opposing sentiment classes, such as "gun control" in bin 1 (Table 6.12), "abortion", "evolution", "gay marriage" and "gun control" in bin 2 (Table 6.13) and "gun control" in bin 4 (Table 6.15). An illustration of a topic with opposing sentiment classes is "gun control" in bin 2, mostly related to sentiment classes 4 and 2 (see Figure 6.13). Here one can see the sentiment going from negative to neutral in class 2 and from neutral to negative in class 4. Excerpts from two discussions from the respective classes are shown in Examples 6.3.2 and 6.3.3. In Example 6.3.2 the discussion starts with authors expressing low sentiments, using terms like "damn", "nonsense", and "pretty obvious", but the discussion progresses towards a more neutral/positive sentiment, where author 102 even expresses that a previous author explained their point well.

Example 6.3.2: Discussion from bin 2, class 2 where sentiment gets higher

...

Message 2 (author 148): Make a damn point already. Posting nonsense like this proves nothing, other than the fact that you're an internet troll. (Sentiment score: -**0.23**)

Message 3 (author 437): Isn't the point pretty obvious- that exposure to a gun can raise testosterone levels and trigger violent and aggressive thoughts and actions? (Sentiment score: -**0.86**)

Message 4 (author 148): Not with you Mr. Muddy Waters. It's a bunch of bull****, another attempt at a causation/correlation bait and switch tactic without any evidence to back it up. There are plenty of people who have constant access to firearms, and yet never display violent tendencies. You really should be careful about posting nonsense like this. If your article has any validity, it would be grounds for disarming the police, since them having access to guns would make them more violent towards the people, and more likely to kill innocent bystanders. (Sentiment score: -**0.04**)

...

Message 25 (author 102): Perhaps I hadn't been following closely enough previously to understand your position. I felt you explained your position well so didn't feel a need to pursue things further. (Sentiment score: **0.14**)

Message 26 (author 37): Cheers mate.emoticonXHoho (Sentiment score: **0.24**)

...

In Example 6.3.3, the opposite pattern is shown, where the discussion starts off with authors expressing neutral sentiments, but the discussion progresses to authors expressing more negative sentiments using terms like "Ohh give it a rest...", "YOU", "nazi dirty work".

Example 6.3.3: Discussion from bin 2, class 4 where sentiment gets lower

...

Message 2 (author 28): I cite from many sources. I don't make the mistake of just depending on one source. (Sentiment score: **0.13**)

Message 3 (author 20):From the top of my head, you have used: 1) Sarah Brady 2) NEJM 3) Sarah Brady 4) and Sarah Brady. Was it you who brought up assault rifles and the Department of Justice? The source that clearly states that assault rifles only helped police in uniform? It was a while back. Please help my memory and detail your sources. (Sentiment score: **0.09**)

...

Message 30 (author 415): Oh give it a rest... Just because you have the attention span of a lemur does not mean I am posting "propaganda" Kelvin will vouch for the depth I have contributed in the past... So if you are so willing to violate my civil rights will YOU be coming to get my guns or will you be sending someone elses child to do your nazi dirty work? (Sentiment score: -**0.50**)

Message 31 (author 17): Talk about begging the question! I don't want your gun, and if such a law were passed it's not my job to enforce the law. (Sentiment score: **0.10**)

Message 32 (author 415): I see you are willing to violate my constitutional rights yet you expect someone else to do your dirty work.... How typical. (Sentiment score: -**0.73**)

...

Inspecting the results the other way around, we see that some classes highly fit one topic. In bin 1 (Table 6.12), class 6 is almost entirely represented by the topic "gun control", in bin 2 class 2 and 4 mostly by "gun control" (Table 6.13), in bin 3 class 5 and 7 (Table 6.14), etc. In each bin, such classes can be found for the topic "gun control". These have been plotted in Appendix G.1. The sentiment trends live mostly between -0.25 (slightly negative) and 0 (neutral), where some sentiment classes increase and some decrease.

Summarizing the results across the bins of discussion lengths, discussions about the topic "existence of God" mostly grow more negative. Some topics, such as "abortion" have many variations in the corresponding sentiment classes, and specifically "gun control" has two types of sentiment classes, one growing more negative and one growing more positive. Some other topics show patterns across a few bins, but not the rest. An example is "evolution", which corresponds to classes with low affective shifts in half of the bins, but in the other half its classes are more spread out.

		Topics						Affective shift
		abortion	climate change	evolution	existence of God	gay marriage	gun control	
Sentiment classes	4	10	1	20	4	4	10	-0.12284
	6	1	0	0	0	1	36	-0.08653
	1	14	2	31	0	7	11	-0.05483
	3	6	2	23	2	7	3	0.127225
	2	9	3	12	1	4	23	0.177408
	5	6	0	1	0	0	14	0.228101

Table 6.12: Contingency table showing the interrelation between sentiment classes and topics, bin 1 (7-22 posts)

		Topics						Affective shift
		abortion	climate change	evolution	existence of God	gay marriage	gun control	
Sentiment classes	5	17	4	37	3	8	8	-0.1039
	4	7	1	3	1	5	36	-0.10302
	3	5	1	16	5	5	3	0.042708
	1	13	5	22	1	16	21	0.115959
	2	7	0	2	1	2	28	0.203079

Table 6.13: Contingency table showing the interrelation between sentiment classes and topics, bin 2 (23-33 posts)

		Topics							Affective shift
		abortion	climate change	communism vs capitalism	evolution	existence of God	gay marriage	gun control	
Sentiment classes	1	8	3	1	30	7	6	6	-0.10933
	2	10	0	0	4	0	2	18	-0.07861
	6	4	1	2	21	5	9	1	-0.04148
	7	3	1	0	0	1	0	17	0.048639
	4	10	3	3	31	2	9	3	0.064126
	3	13	0	2	6	2	3	12	0.202685
	5	8	0	0	1	0	0	26	0.269407

Table 6.14: Contingency table showing the interrelation between sentiment classes and topics, bin 3 (34-50 posts)

		Topics					Affective shift
		abortion	*evolution*	*existence of God*	*gay marriage*	*gun control*	
Sentiment classes	*6*	8	45	10	13	1	-0.06561
	2	32	26	3	6	31	-0.00523
	4	6	0	0	2	25	0.159404
	3	14	28	4	13	8	0.169169
	5	12	1	0	1	37	0.224891
	1	10	2	2	6	5	0.25188

Table 6.15: Contingency table showing the interrelation between sentiment classes and topics, bin 4 (51-86 posts)

		Topics					Affective shift
		abortion	*evolution*	*existence of God*	*gay marriage*	*gun control*	
Sentiment classes	*5*	7	17	3	6	1	-0.15862
	1	3	7	8	0	0	-0.13856
	7	8	15	1	8	3	0.055734
	3	3	0	0	0	21	0.069402
	2	11	11	1	2	5	0.104009
	4	9	1	0	2	4	0.113944
	8	10	2	0	1	5	0.133831
	6	9	0	0	2	10	0.233061

Table 6.16: Contingency table showing the interrelation between sentiment classes and topics, bin 5 (87-136 posts)

		Topics				Affective shift
		abortion	*evolution*	*gay marriage*	*gun control*	
Sentiment classes	*4*	2	6	4	0	-0.13589
	6	1	0	0	4	-0.08988
	3	7	0	1	5	0.004629
	5	8	2	2	2	0.078875
	2	1	8	1	0	0.144985
	1	1	0	1	6	0.297916

Table 6.17: Contingency table showing the interrelation between sentiment classes and topics, bin 6 (137-186 posts)

6.3.5 Patterns across alignment, sentiment, and topic

Comparing the contingency tables per topic of alignment vs. topic, and the contingency tables of sentiment vs. topic does not show many similarities. One pattern that can be seen is that "gun control" is both related to opposing types of alignment classes (corresponding both to classes with low and high alignment changes) and opposing types of sentiment classes (with both low and high affective shifts). For the other topics, no obvious patterns showed.

6.4 Discussion

The results imply several things. First of all, the Pearson correlation coefficient for average sentiment and alignment showed that there was no linear dependency between them. This implies that there is no relation, or that it is far more complicated than we have tried to measure. We saw a slightly higher correlation between the average alignment and the minimum and maximum sentiment, but we argue that this could be caused by the discussion length below.

The results show a high correlation between the average alignment and the discussion length. However, this was expected as our definition of alignment, the time-based overlap, depends on the discussion length as vocabularies increase with longer discussions. The results also show a correlation between sentiment and length, but that can also be explained. When a discussion contains more posts, there are more possibilities for extreme sentiments to be expressed. Longer discussions indeed showed to have a lower minimum sentiment score and a higher maximum sentiment score. If we look at the average sentiment score compared to the discussion length, these extremes cancel each other out.

These correlations could cause the slightly higher correlation between the average alignment and the minimum and maximum sentiment, as longer discussions both have a more extreme sentiment and a higher alignment. This is reflected by the joint distributions of the average alignment and the minimum and maximum sentiment respectively, as they show that as discussions have a higher time-based overlap, the minimum sentiment score gets lower and the maximum sentiment score gets higher.

Another implication of the results is that there is no association between the alignment classes and sentiment classes. Again, this can imply two things: that there is no relation at all, or that the relation is far more complex than we have tried to measure. We hypothesize it is the latter option, as we have found a different interesting association, which will be discussed below.

We saw a low association between alignment and topic, implying that either topics are slightly associated with alignment, or perhaps that some topics are more associated with alignment. Some topics, such as "evolution" are represented by clear classes. Other classes, like "gun control" and "climate change" are clearly represented by two separate trends, one more flat and one more steep. Looking at it from another perspective, how classes are formed by topics, we could not find patterns.

The results showed some clear interplay between sentiment patterns and topic. We saw some highly represented classes per topic (e.g. "existence of God" and "gay marriage"), and also topics per class (e.g. "gun control"). Not all topics were however represented by certain sentiment classes: some topics were spread over multiple classes (e.g. "evolution"), and some were split over opposing classes (e.g. "gun control").

This implies that people behave differently in expressing sentiments on different topics. Furthermore, it sparks the question of whether the topic of a discussion might be a factor influencing the interplay between sentiment and alignment, or whether sentiment and alignment are independent. This cannot be answered with the current research and might be an interesting direction for future work.

We have seen some limitations during the interplay analysis. One is that not all discussions in the dataset were annotated with their topic (see Section 6.1). This means that there was less data available for the topic analyses. Higher bins had a lower amount of annotated discussions, which could mean that any patterns we could induce could be overfitted, making the results less reliable and representable. Bin 7 (discussions with 187-236 posts) and bin 8 (discussions with 237-286 posts) had too few annotated discussions (<50) to be examined entirely. Future work could investigate the topic of longer discussions.

A related improvement could be made in the methods. Currently, we have first clustered the discussions based on alignment and sentiment trends, and only then removed discussions that were not annotated by topic to investigate their association with the topic. This means that the clustering includes discussions that we do not know the topic of. It could be that these discussions have different topics, have less clearly defined topics, or perhaps even discuss multiple topics. We do not know their influence on the trends. Future work could extract the discussions annotated with their topic and cluster these, to explore if that gives clearer results.

We have seen that investigating the average alignment and average sentiment was not very useful from the correlations. Future work could look into aggregating messages in discussions in another way than taking the average.

A returning limitation is the choice for the measure of alignment (see also Chapter 4). We could see how our presented measure is dependent on the discussion length, which could also be affecting our results on the interplay between sentiment and alignment. Future work could use a different measure and investigate whether that returns different results. It could also look at different alignment levels (such as syntax or the semantic level), as they might generally be less dependent on discussion length.

One of the topics that stood out in the interplay analysis was "gun control", having two separate types of trends for both alignment and sentiment and being a topic that highly fits certain alignment and sentiment classes. It might be interesting for future work to just investigate discussions about this topic further.

6.5 Conclusion

With the results of the previous chapters, heat plots were plotted and the Pearson correlation coefficient was computed for the (average) alignment and the (average, minimum, and maximum) sentiment in discussions. With these results, **Q3.1** can now be answered: we did not find a relation between sentiment and alignment. The average sentiment & alignment shows no correlation, but there was a slight correlation between alignment and the minimum & maximum sentiment: more alignment seems to be related to more extreme sentiments. However (as discussed before), this could be caused by the discussion length. So, if there is an interplay, it is more complicated than just a linear correlation between these two variables.

Next, heat plots and the Pearson correlations were also computed for the (average) alignment and (average, minimum, and maximum) sentiment against the discussion length. With those results, **Q3.2** can be answered. We found that longer discussions have a higher average alignment, but (as discussed before) this is by the definition of our alignment measure. We also found that longer discussions have more extreme sentiments, but that was also expected as there are more possibilities for extreme posts to occur. To conclude, we learned that the average is not an insightful measure for sentiment, as the extremes have just been canceled out and were averaged to neutral.

From the previous chapters, trends of alignment (alignment classes) and trends for sentiment (sentiment classes) were extracted. For each bin of discussion lengths, we set the sentiment classes against the alignment classes and created contingency tables, and computed Cramér's V to find their association. With these results, **Q3.3** can be answered. We have not found a relationship between sentiment and alignment over time. This could either mean that it does not exist, or it could mean that it is far more complex than expected.

In the final step, we set out the alignment and sentiment trends against the topic in discussions, created more contingency tables, and computed Cramér's V. With that, **Q3.4** can be answered. We found some interplay between alignment over time and topic, though patterns are not always clear. We found a higher association between sentiment over time and topic. Some topics (such as "existence of God") correspond to certain sentiment classes, some (such as "abortion") are more spread between classes, and some have two distinct patterns, such as "gun control" with sentiment either becoming more negative, or more neutral. We also saw that for quite some classes, the topic "gun control" fitted best. Looking at these classes, we found that the sentiment in these classes ranged between slightly negative and neutral.

Chapter 7

Discussion

The intermediate results have already been discussed in Sections 4.5, 5.5, and 6.4. In this chapter, we discuss the research from a higher level.

As discussed before, the results implied that (finding) a relationship between sentiment and alignment is far more complex than we initially thought. We proposed that more factors than just the two might be involved, as we have shown that there was some association between alignment and topic, and even more between sentiment and topic. It appeared that the length of the discussion might be involved as well, and there might be other factors that we do not yet know about.

These results do not confirm what Niederhoffer and Pennebaker proposed, which was that a lower sentiment would correlate to higher alignment (Niederhoffer & Pennebaker, 2002). They also do not confirm that alignment can be used to achieve higher sentiments as hypothesized by Bernhold and Giles (Bernhold & Giles, 2020). However, this study also does not necessarily rule out their ideas. Since this study was the first to thoroughly investigate this relation, future work may have more conclusive findings.

The subparts of this research also contribute to the understanding of human behavior. The clustering results are particularly fascinating: out of a lot of data points, the clustering picked out conversation typologies; implying different types of trends exist in both alignment and sentiment.

We encountered various challenges, both in resources and methods. A challenge with forum data is that much is unknown about the behavior of users. It is therefore unclear what the impact of previous posts is. One can not be sure which posts people read before replying, sometimes which posts authors wanted to reply to, which posts influence authors most, in what order authors read posts, etc. The 4forums dataset poses another challenge, as 4forums provided different UI options to users (linear, threaded, mixed), and it is unknown which UI mode users used.

Another limitation of the resources was that no recent discussion datasets were available; the one currently used is rather old. Twitter data might have been interesting, but it recently changed its policies. Reddit data was available but was decided against, as posts on Reddit are longer which might be less interesting for alignment, as the threads are less like discussions and more like essays.

A limitation of the methods is that the results depend a lot on the measures used. For alignment, many levels of alignment could have been investigated, and there are many measures available per level. However, few of them measure alignment over time in multiparty discussions. Furthermore, many measures were not described properly, skipping over details which made them difficult to follow and implement. Other measures took too much computational power or time to run. Therefore, a simple new measure was presented, but it still has room for improvements, for instance making it less dependent on discussion length. This work can thus become the baseline for measuring alignment over time, on which other research can continue to improve.

Another limitation of the methods is that averaging alignment and sentiment proved not to be very informative, especially in the sentiment distribution. This was also discussed in Section 5.5. Future work could investigate more into the minimum and maximum sentiment scores, or perhaps find some other solution than averaging.

There are many directions that future work could go in, of which some have already been discussed in the previous chapters. An additional idea for future work is to investigate other factors that might be related to (either or both) sentiment and alignment. The initial idea of this research was to also investigate agreement, but the dataset did not allow for analyzing agreement over time. However, the agreement could very well be another factor in the interplay, as Van der Pol et al. have already found that interlocutors align more if they disagree (van der Pol et al., 2016), and one could hypothesize that people might express more positive sentiment if they agree and more negative if they disagree. Other factors that might be interesting to investigate are the original stance of the interlocutors, their authority (power), their engagement, their open-mindedness, the number of people in a discussion, what kind of social groups are mixed, and how serious or casual a discussion is.

Currently, posts are structured in a discrete order, without taking into account real-time. It could be interesting to look at the actual time instead, including gaps in between, as this might lead to different trends and results. The dataset allows for such an addition as it contains timestamps.

Furthermore, we have looked at which classes of sentiment and alignment could be found with a clustering algorithm, and how they are associated with each other and with the topic of discussions. Aside from knowing if they are related, it would be even more interesting to see in which manner they are related, such as specific influences on each other (would they positively reinforce each other, or do they contrast?), and in what order and which manner that presents itself (does one cause the other?).

Though this research has sparked a lot of new questions and suggestions for future work, concrete next steps could be the following:

- Improve the resources: create and use a more modern, complete dataset that includes modern language (about contemporary topics), emojis, and perhaps longer discussions, optimally with more (and all posts in) discussions being annotated with interesting factors such as topic, agreement and others which have been mentioned before.

- Improve the methods: update the time-based overlap or find a different measure of alignment (perhaps for different levels), that is not as dependent on discussion length. As discussed in Chapter 4, time-based overlap could for instance be improved by adding a penalty to terms used in older posts, essentially adding a decay cost. Furthermore, one could look at different trends of different authors, tackling multiparty discussions in a more detailed way.

- Investigate related topics: with improved resources and methods, one could further investigate the interplay with other factors that might be involved, such as the agreement, (the difference in) original stance of the interlocutors on the topic, power, engagement, etc.

Chapter 8

Conclusion

In this research, we worked towards answering the following main question (presented in Chapter 1):

> *How does linguistic alignment relate to the expressed sentiments of interlocutors in forum posts about political topics?*

We answered this question by first posing smaller sub-questions. Part 1 investigated the linguistic alignment, which was discussed in Chapter 4. To investigate lexical alignment, we created a new measure: the time-based overlap. The time-based overlap measures the amount of overlap between a post and previous posts by other authors. We applied it to all messages in all discussions and found the alignment distribution. With that, we answered **Q1.1**: the alignment distribution is bell-shaped in online discussions on 4forums, with a mean lexical alignment of 0.6, and it ranges between 0.2 and 0.9.

Next, the discussions were divided into bins based on their number of posts. Trends of the time-based overlap for each discussion were extracted and a moving average was applied to smoothen them. Then we applied k-means clustering for time-series data. With these clustering results, **Q1.2** was answered: the clustering found distinct classes. The alignment in terms of time-based overlap mostly increases with a curved shape, stabilizing at some level of alignment. In a few classes, the trend goes down towards the end. The trends mostly differ in at which alignment they start, how steep the increase is, and how high the alignment reaches.

Part 2 investigated the sentiment expressed in discussions, which was described in Chapter 5. We computed the compound, minimum, and maximum sentiment score of all messages in discussions using VADER, and computed the minimum, maximum, and average sentiment score per discussion. With that, we computed percentiles and plotted histograms. This answered **Q2.1**: the average sentiment is bell-shaped with a mean around a neutral sentiment. However, the minimum and maximum sentiment distributions show a big range of sentiments being present within discussions.

Then, we divided discussions into the same bins as for the alignment analysis and applied a moving average to the compound sentiment scores, after which we applied clustering. With these results, we answered **Q2.2**: distinct classes of sentiment trends were found per bin of discussion lengths, changing from slightly towards one polarity to the other one, but also going back and forth, mostly lying around a neutral sentiment. The trends never reach the thresholds of positive polarity (0.5) and negative polarity (-0.5) and their changes are not large.

Part 3 investigated the interplay between alignment and sentiment, and additionally, discussion length and topic, which was described in Chapter 6. To find the interplay, the results of the previous parts (Chapter 4 and 5) were used. The Pearson correlation coefficient was used to find the correlation between alignment, sentiment, and discussion length. With that, we can answer **Q3.1**: we found no correlation between the average sentiment and alignment, a slight correlation between the alignment and the minimum & maximum sentiment. A higher alignment appears to be related to more extreme sentiments (in both directions of the polarities), however, this is most likely caused by the influence of the discussion length. So, the interplay between sentiment and alignment as seen in discussions overall is more complicated than just a linear correlation. The answer to **Q3.2** is that we found that longer discussions have higher alignment, but that is caused by the definition of the alignment measure. We also found that longer discussions have more extreme sentiments, but this was also expected as there are more possibilities for extreme posts to occur in longer discussions than in shorter ones. We learned that taking the average is not an informative measure as the extremes were just averaged out to neutral.

Next, we created contingency tables (per bin of discussion lengths) of sentiment classes against alignment classes (which followed from the clustering results), and used them to compute Cramér's V to find the association between the two. With these results, **Q3.3** was answered: we have not found a relation between the sentiment and alignment over time. This could either mean that the relation does not exist, or it could mean that it is much more complex than expected, where other factors influence the interplay.

We also created contingency tables (per bin of discussion lengths) of the discussions that were annotated with their topic, setting alignment and sentiment against the topic respectively. Again, Cramér's V was computed. With that, we answered **Q3.4**: we found some interplay between alignment over time and topic, though the association is low. We found a higher association between the sentiment over time and the topic. Some topics clearly correspond to certain sentiment classes, some topics are more spread between classes, and some clearly correspond to two distinct patterns.

Finally, the main research question can be answered. We did not find an interplay between lexical alignment and sentiment. However, an association between sentiment and the discussion topic was found, and our results indicate that the discussion length might also be related. Therefore, if there is an interplay between sentiment and linguistic alignment, it is far more complex than a simple correlation and likely depends on other factors such as topic and discussion length.

References

Abbott, R., Ecker, B., Anand, P., & Walker, M. A. (2016). Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4445–4452.

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015, October). Time-series clustering – A decade review. *Information Systems*, *53*, 16–38. doi: 10.1016/j.is.2015.04.007

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013, October). How Noisy Social Media Text, How Diffrnt Social Media Sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 356–364). Nagoya, Japan: Asian Federation of Natural Language Processing.

Bernhold, Q. S., & Giles, H. (2020, March). Vocal Accommodation and Mimicry. *Journal of Nonverbal Behavior*, *44*(1), 41–62. doi: 10.1007/s10919-019-00317-y

Bhagat, M., & Bakariya, B. (2022). Sentiment Analysis Through Machine Learning: A Review. In G. Mathur, M. Bundele, M. Lalwani, & M. Paprzycki (Eds.), *Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications* (pp. 633–647). Singapore: Springer Nature. doi: 10.1007/978-981-16-6332-1_52

Brinberg, M., & Ram, N. (2021, July). Do New Romantic Couples Use More Similar Language Over Time? Evidence from Intensive Longitudinal Text Messages. *Journal of Communication*, *71*(3), 454–477. doi: 10.1093/joc/jqab012

Carrick, T., Rashid, A., & Taylor, P. J. (2016, August). Mimicry in online conversations: An exploratory study of linguistic analysis techniques. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 732–736). San Francisco, CA, USA: IEEE. doi: 10.1109/ASONAM.2016.7752318

Condon, W. S., & Ogston, W. D. (1967, September). A segmentation of behavior. *Journal of Psychiatric Research*, *5*(3), 221–235. doi: 10.1016/0022-3956(67)90004-0

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web - WWW '11* (p. 745). Hyderabad, India: ACM Press. doi: 10.1145/1963405.1963509

Danescu-Niculescu-Mizil, C., & Lee, L. (2011, June). Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 76–87). Portland, Oregon, USA: Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012, April). Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web* (pp. 699–708). Lyon France: ACM. doi: 10.1145/2187836.2187931

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv. (arXiv:1810.04805 [cs])

Doyle, G., & Frank, M. C. (2016, August). Investigating the Sources of Linguistic Alignment in Conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 526–536). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/P16-1050

Doyle, G., Goldberg, A., Srivastava, S., & Frank, M. (2017). Alignment at Work: Using Language to Distinguish the Internalization and Self-Regulation Components of Cultural Fit in Organizations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 603–612). Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/P17-1056

Doyle, G., Yurovsky, D., & Frank, M. C. (2016, April). A Robust Framework for Estimating Linguistic Alignment in Twitter Conversations. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 637–648). Montréal Québec Canada: International World Wide Web Conferences Steering Committee. doi: 10.1145/2872427.2883091

Dubuisson Duplessis, G., Clavel, C., & Landragin, F. (2017). Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 71–81). Saarbrücken, Germany: Association for Computational Linguistics. doi: 10.18653/v1/W17-5510

Dubuisson Duplessis, G., Langlet, C., Clavel, C., & Landragin, F. (2021, June). Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. *Language Resources and Evaluation*, *55*(2), 353–388. doi: 10.1007/s10579-021-09532-w

Duran, N. D., Paxton, A., & Fusaroli, R. (2019, August). ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychological Methods*, *24*(4), 419–438. doi: 10.1037/met0000206

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012, August). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, *23*(8), 931–939. (Publisher: SAGE Publications Inc) doi: 10.1177/0956797612436816

Giachanou, A., & Crestani, F. (2017, June). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys*, *49*(2), 1–41. doi: 10.1145/2938640

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics.* Cambridge: Cambridge University Press.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant

Supervision. *CS224N project report, Stanford*, *1*(12).
Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*.
Grover, V. (2022). Exploiting Emojis in Sentiment Analysis: A Survey. *Journal of The Institution of Engineers (India): Series B*, *103*(1), 259–272. doi: 10.1007/s40031-021-00620-7
Healey, P. G. T., Purver, M., & Howes, C. (2014, June). Divergence in Dialogue. *PLoS ONE*, *9*(6), e98598. doi: 10.1371/journal.pone.0098598
Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* Seattle, Washington: ACM.
Hung, L., & Alias, S. (2023). Beyond Sentiment Analysis: A Review of Recent Trends in Text Based Sentiment Analysis and Emotion Detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *27*(1), 84–95. doi: 10.20965/jaciii.2023.p0084
Hutto, C. J., & Gilbert, E. (2014, May). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. (Number: 1) doi: 10.1609/icwsm.v8i1.14550
Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue* (No. 8). New York: Academic Press.
Jones, S., Cotterill, R., Dewdney, N., Muir, K., & Joinson, A. (2014, August). Finding Zelig in Text: A Measure for Normalising Linguistic Accommodation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 455–465). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.
Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (Third edition draft ed.).
Kilgarriff, A. (2001, December). Comparing Corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133. doi: 10.1075/ijcl.6.1.05kil
Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 538–541. (Number: 1) doi: 10.1609/icwsm.v5i1.14185
Kuzilek, J., Kravcik, M., & Sinha, R. (2020, June). Analysis of Discussion Forum Data as a Basis for Mentoring Support. *Joint Proceedings of Workshops IMHE 2020 and WELL4SD 2020*.
Levallois, C. (2013, June). Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 414–417). Atlanta, Georgia, USA: Association for Computational Linguistics.
Lexalytics. (2015). *Sentiment extraction - measuring the emotional tone of content* (Tech.

Rep.). Author.

Liu, B. (2010, January). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*, 627–666.

Macqueen, J. (1967, June). Some Methods For Classification And Analysis Of Multivariate Observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probabilit*, *1*(14), 218–297.

Medhat, W., Hassan, A., & Korashy, H. (2014, December). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. doi: 10.1016/j.asej.2014.04.011

Müller, M. (2007). Dynamic Time Warping. In *Information Retrieval for Music and Motion* (pp. 69–84). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-74048-3_4

Niederhoffer, K. G., & Pennebaker, J. W. (2002, December). Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360. (Publisher: SAGE Publications Inc) doi: 10.1177/026192702237953

Nielsen, F. (2011, March). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.* arXiv. (arXiv:1103.2903 [cs])

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010, May). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 122–129. (Number: 1) doi: 10.1609/icwsm.v4i1.14031

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015* (Tech. Rep.). Austin: University of Texas at Austin.

Pickering, M. J., & Garrod, S. (2004, April). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). doi: 10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2006, October). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, *4*(2-3), 203–228. doi: 10.1007/s11168-006-9004-0

Pilehvar, M. T., Jurgens, D., & Navigli, R. (2013, August). Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1341–1351). Sofia, Bulgaria: Association for Computational Linguistics.

Pinto, H. L., & Rocio, V. (2019). Combining Sentiment Analysis Scores to Improve Accuracy of Polarity Classification in MOOC Posts. In P. Moura Oliveira, P. Novais, & L. P. Reis (Eds.), *Progress in Artificial Intelligence* (pp. 35–46). Cham: Springer International Publishing. doi: 10.1007/978-3-030-30241-2_4

Reitter, D., Keller, F., & Moore, J. D. (2006, June). Computational Modelling of Structural Priming in Dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 121–124). New York City, USA: Association for Computational Linguistics.

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016, December). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 1–29. (Number: 1 Publisher:

SpringerOpen) doi: 10.1140/epjds/s13688-016-0085-1

Services, A. W. (2023). *Amazon Comprehend - Developer Guide.*

Shin, H., & Doyle, G. (2018, June). Alignment, Acceptance, and Rejection of Group Identities in Online Political Discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 1–8). New Orleans, Louisiana, USA: Association for Computational Linguistics. doi: 10.18653/v1/N18-4001

Shobha Rani, M., & Subramanain, S. (2020, October). A Study on Diverse Methods and Performance Measures in Sentiment Analysis. *Recent Patents on Engineering*, *14*. doi: 10.2174/1872212114999201019154954

Soliz, J., & Giles, H. (2014, January). Relational and Identity Processes in Communication: A Contextual and Meta-Analytical Review of Communication Accommodation Theory. *Annals of the International Communication Association*, *38*(1), 107–144. doi: 10.1080/23808985.2014.11679160

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011, June). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. doi: 10.1162/COLI_a_00049

van der Pol, E., Gieske, S., & Fernández, R. (2016, August). Linguistic Style Accommodation in Disagreements. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 120–124). Berlin, Germany: Association for Computational Linguistics. doi: 10.18653/v1/S16-2015

Vergara, S., El-Khouly, M., Tantawi, M. E., Marla, S., & Sri, L. (2017). *Building Cognitive Applications with IBM Watson Services: Volume 7 Natural Language Understanding* (First ed.).

Walker, M. A., Anand, P., Tree, J. E. F., Abbott, R., & King, J. (2012). A Corpus for Research on Deliberation and Debate. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, *12*, 812–817.

Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment Analysis of Twitter Data. *Applied Sciences (Switzerland)*, *12*(22). doi: 10.3390/app122211775

Wang, Y., Reitter, D., & Yen, J. (2014, June). A Model to Qualify the Linguistic Adaptation Phenomenon in Online Conversation Threads: Analyzing Priming Effect in Online Health Community. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics* (pp. 55–62). Baltimore, Maryland, USA: Association for Computational Linguistics. doi: 10.3115/v1/W14-2007

Wang, Y., Reitter, D., & Yen, J. (2017). How Emotional Support and Informational Support Relate to Linguistic Alignment. In D. Lee, Y.-R. Lin, N. Osgood, & R. Thomson (Eds.), *Social, Cultural, and Behavioral Modeling* (Vol. 10354, pp. 25–34). Cham: Springer International Publishing. (Series Title: Lecture Notes in Computer Science) doi: 10.1007/978-3-319-60240-0_3

Ward, A., & Litman, D. (2007). Automatically Measuring Lexical and Acoustic/Prosodic Convergence in Tutorial Dialog Corpora.

Wen, M., Yang, D., & Rosé, C. P. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?

Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 347–354). Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Xu, Y. (2021, February). Global divergence and local convergence of utterance semantic representations in dialogue. In *Proceedings of the Society for Computation in Linguistics 2021* (pp. 116–124). Online: Association for Computational Linguistics.

Xu, Y., & Reitter, D. (2015, June). An Evaluation and Comparison of Linguistic Alignment Measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 58–67). Denver, Colorado: Association for Computational Linguistics. doi: 10.3115/v1/W15-1107

Appendix A

4forums user interface

This appendix contains screenshots of what the website looked like when it was still live.

Figure A.1: What a thread in linear mode looked like in 4forums.com, obtained from the Wayback Machine: https://web.archive.org/web/20120624025816/http://www.4forums.com:80/political/abortion-debates/12680-can-something-legal-immoral-unethical-same-time.html

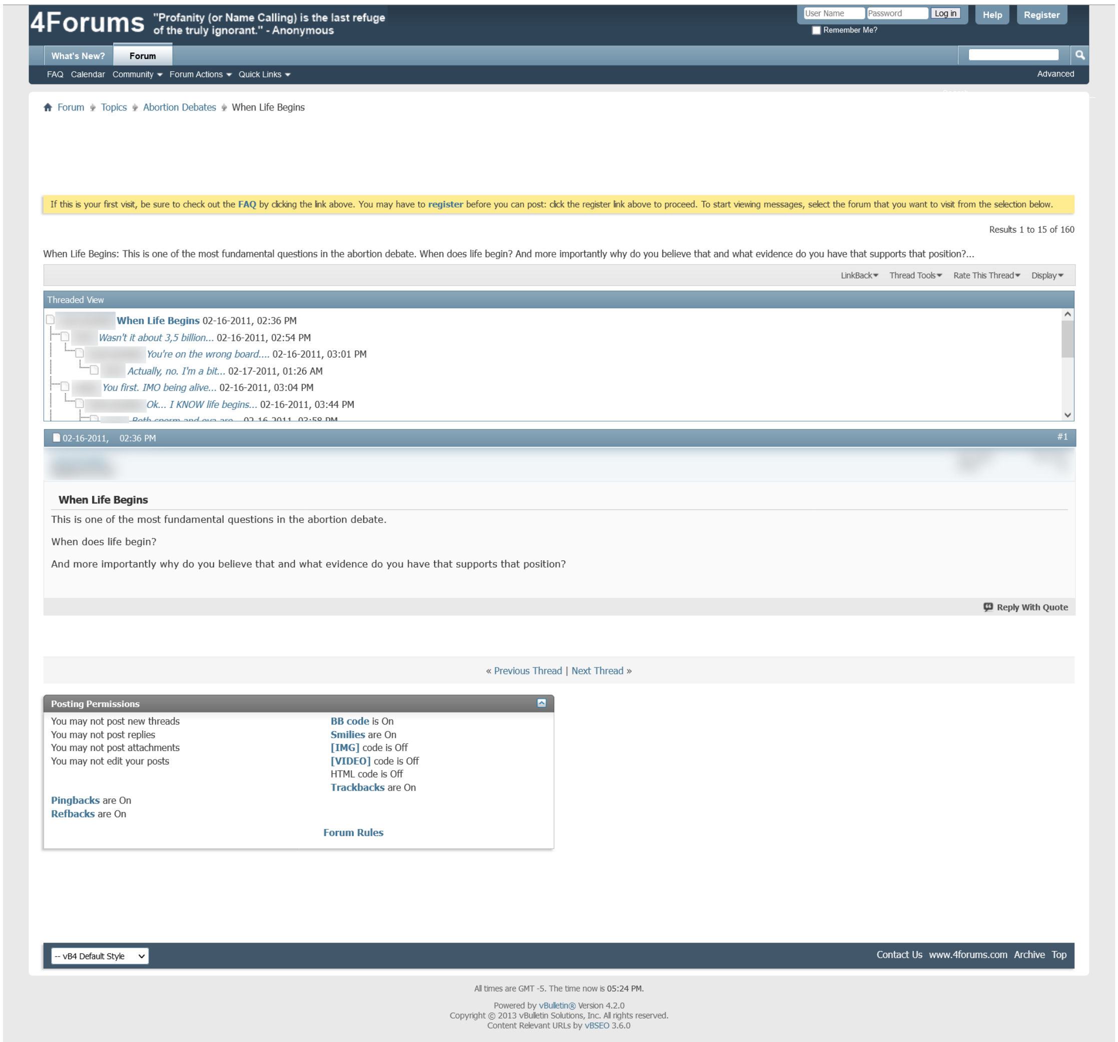


Figure A.2: What a thread in threaded mode looked like in 4forums.com, obtained from the Wayback Machine: `https://web.archive.org/web/20130516222412/http://www.4forums.com/political/abortion-debates/15035-when-life-begins.html`

Appendix B

Additional alignment results

This appendix contains additional (less insightful) results of the time-based overlap analysis.

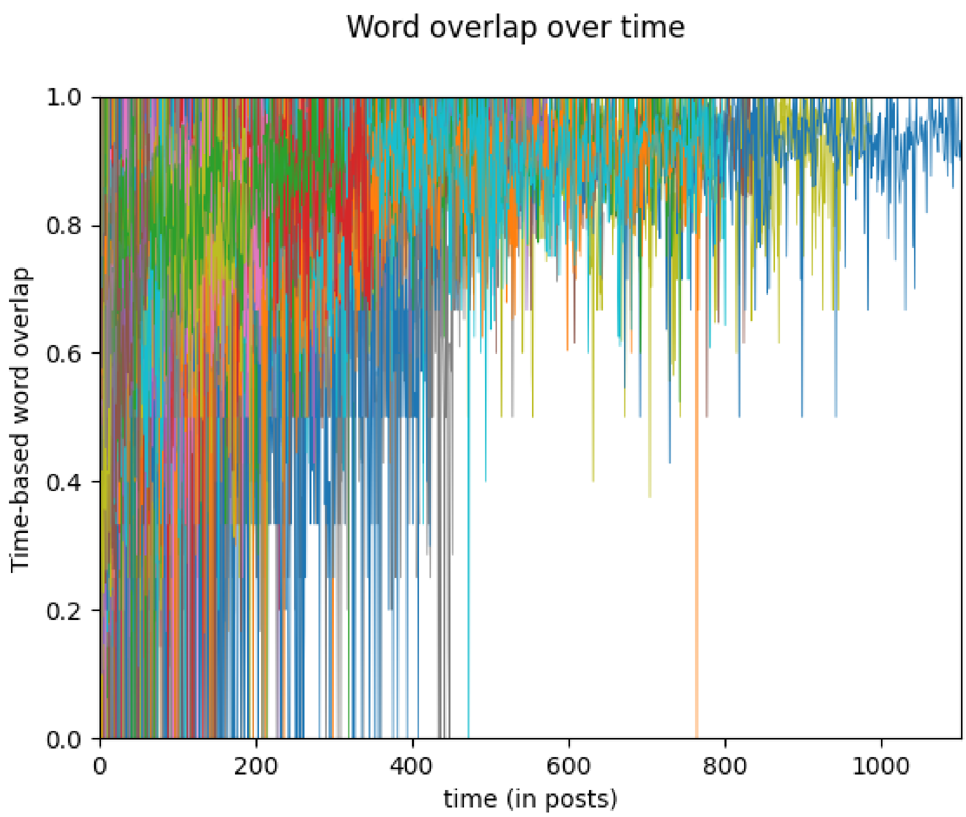


Figure B.1: Time-based overlap over posts for all discussions

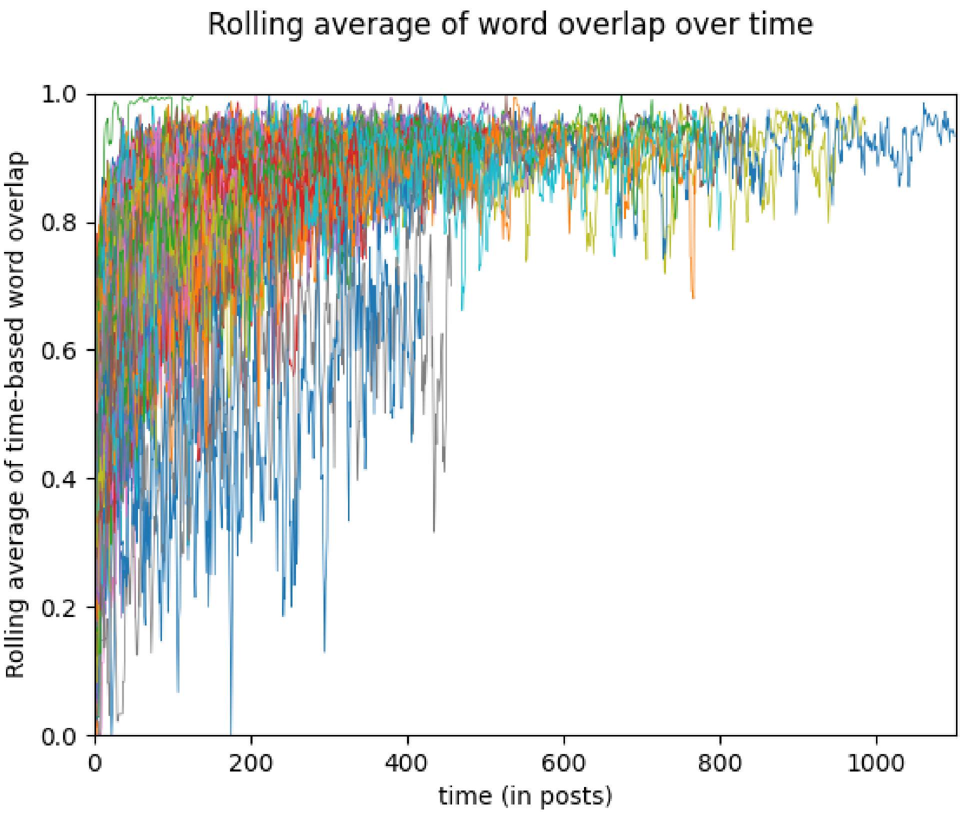


Figure B.2: Rolling averages of time-based overlap over posts for all discussions

Appendix C

Clustering finetuning for alignment

This appendix contains plots that were used to finetune the clustering on time-based overlap.

C.1 Inertia per k for all discussion length bins

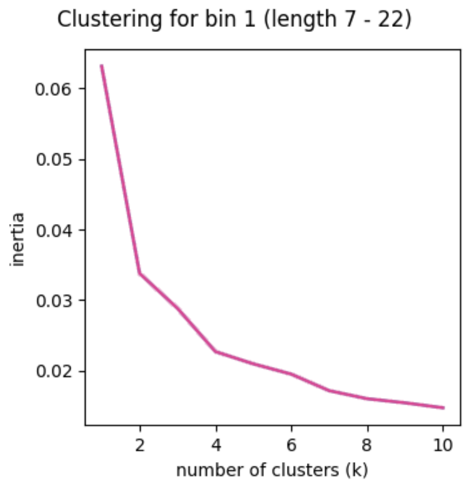


Figure C.1: Inertia per k for the first discussion length bin

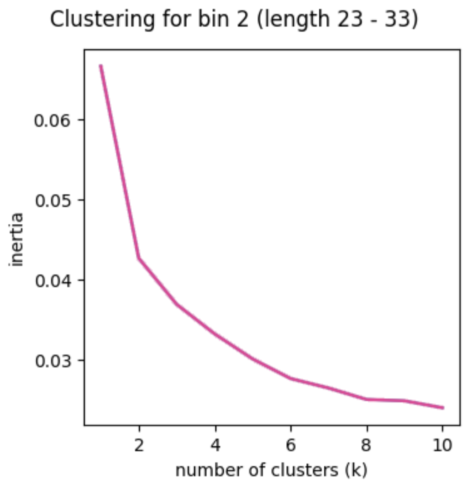


Figure C.2: Inertia per k for the second discussion length bin

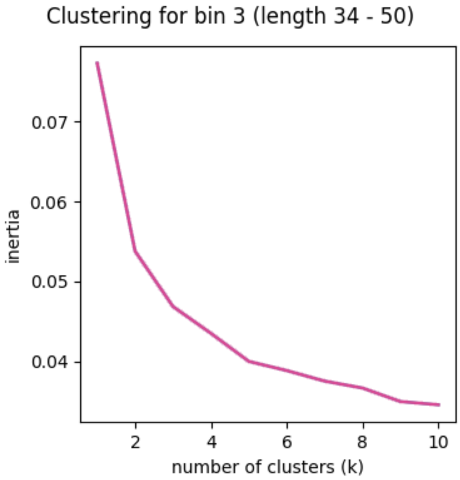


Figure C.3: Inertia per k for the third discussion length bin

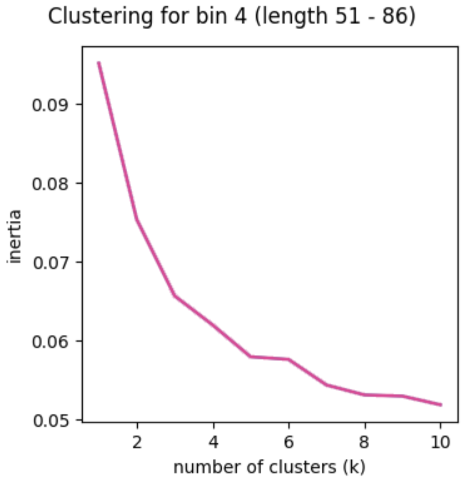


Figure C.4: Inertia per k for the fourth discussion length bin

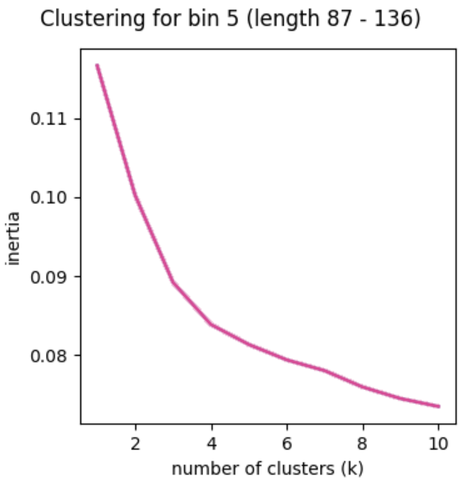


Figure C.5: Inertia per k for the fifth discussion length bin

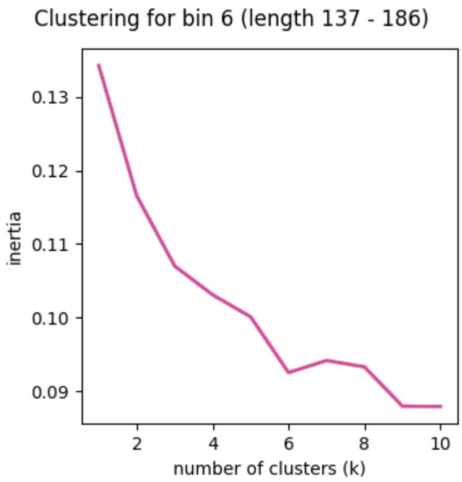


Figure C.6: Inertia per k for the sixth discussion length bin

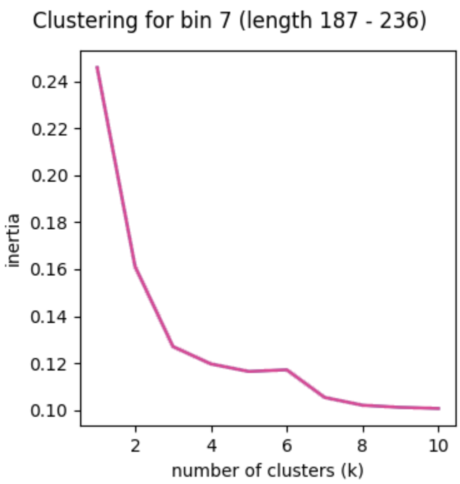


Figure C.7: Inertia per k for the seventh discussion length bin

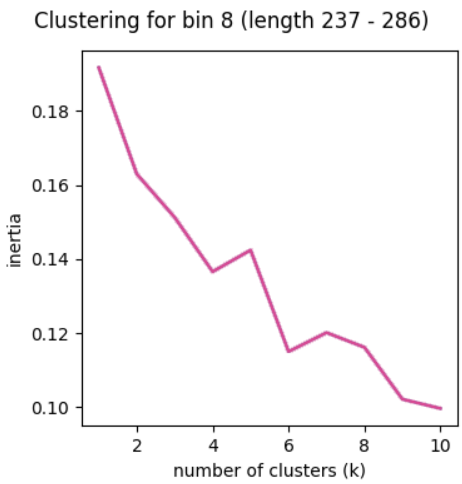


Figure C.8: Inertia per k for the eighth discussion length bin

C.2 Clustering results highlights for bin 6

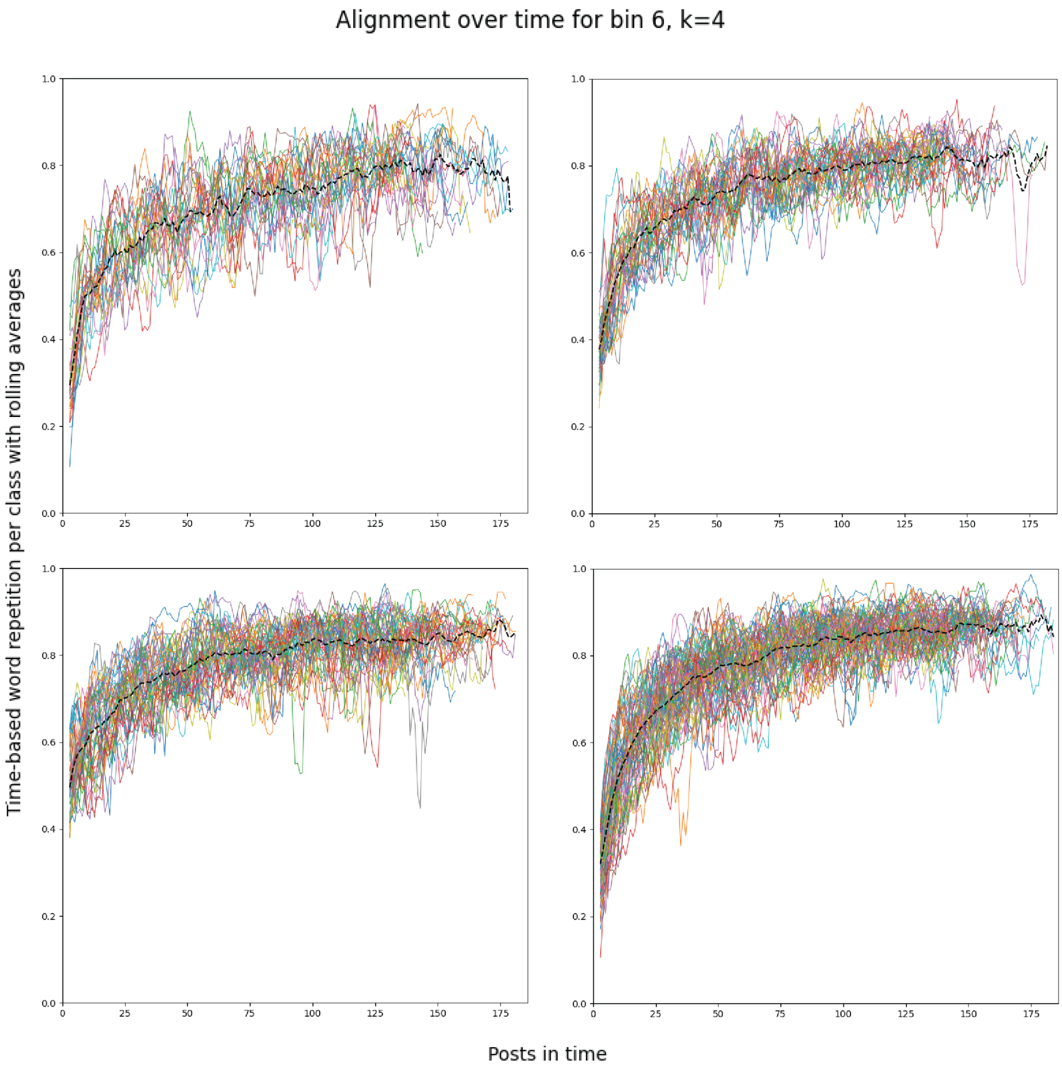


Figure C.9: Clustering results for bin 6, $k = 4$

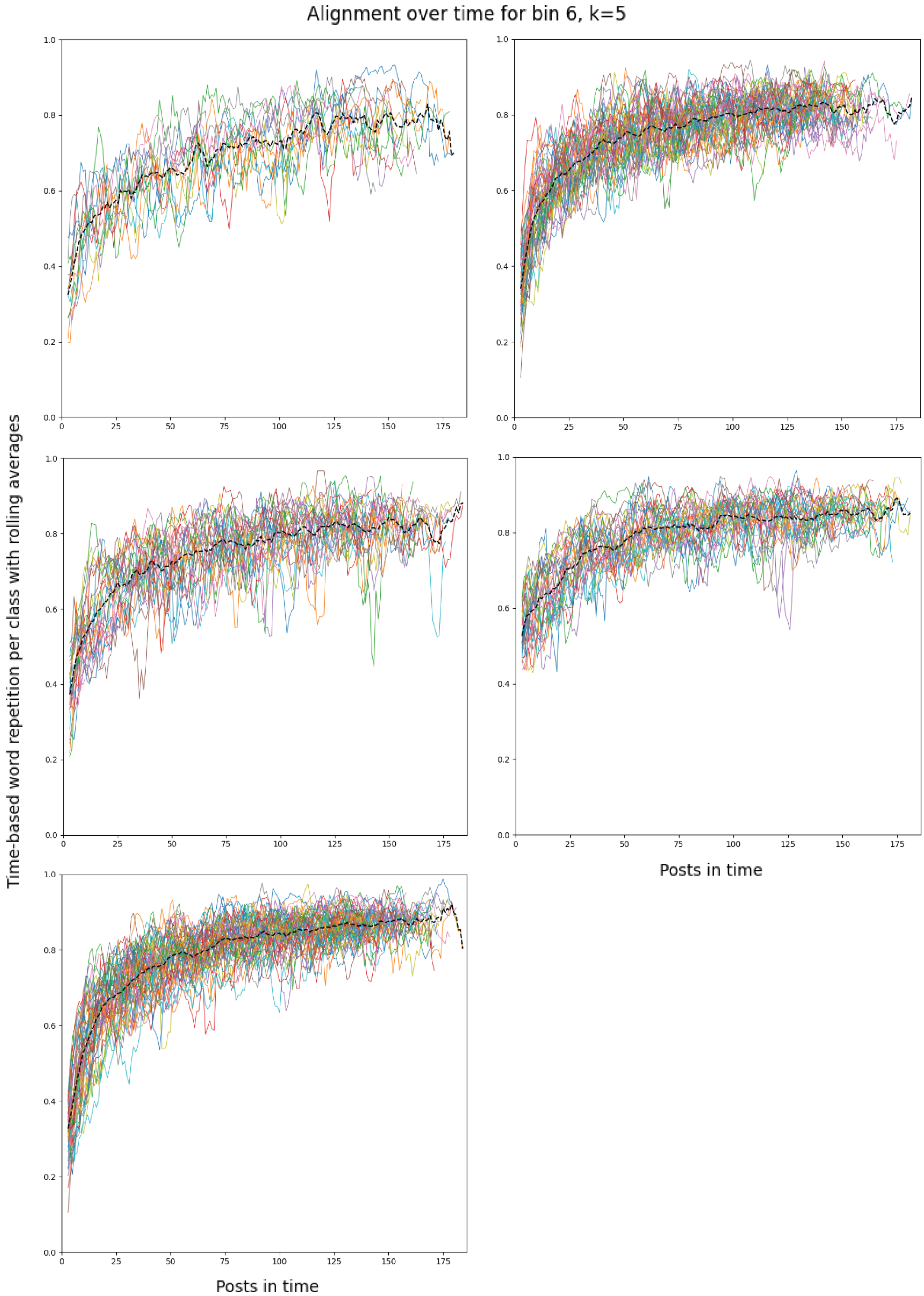


Figure C.10: Clustering results for bin 6, $k = 5$

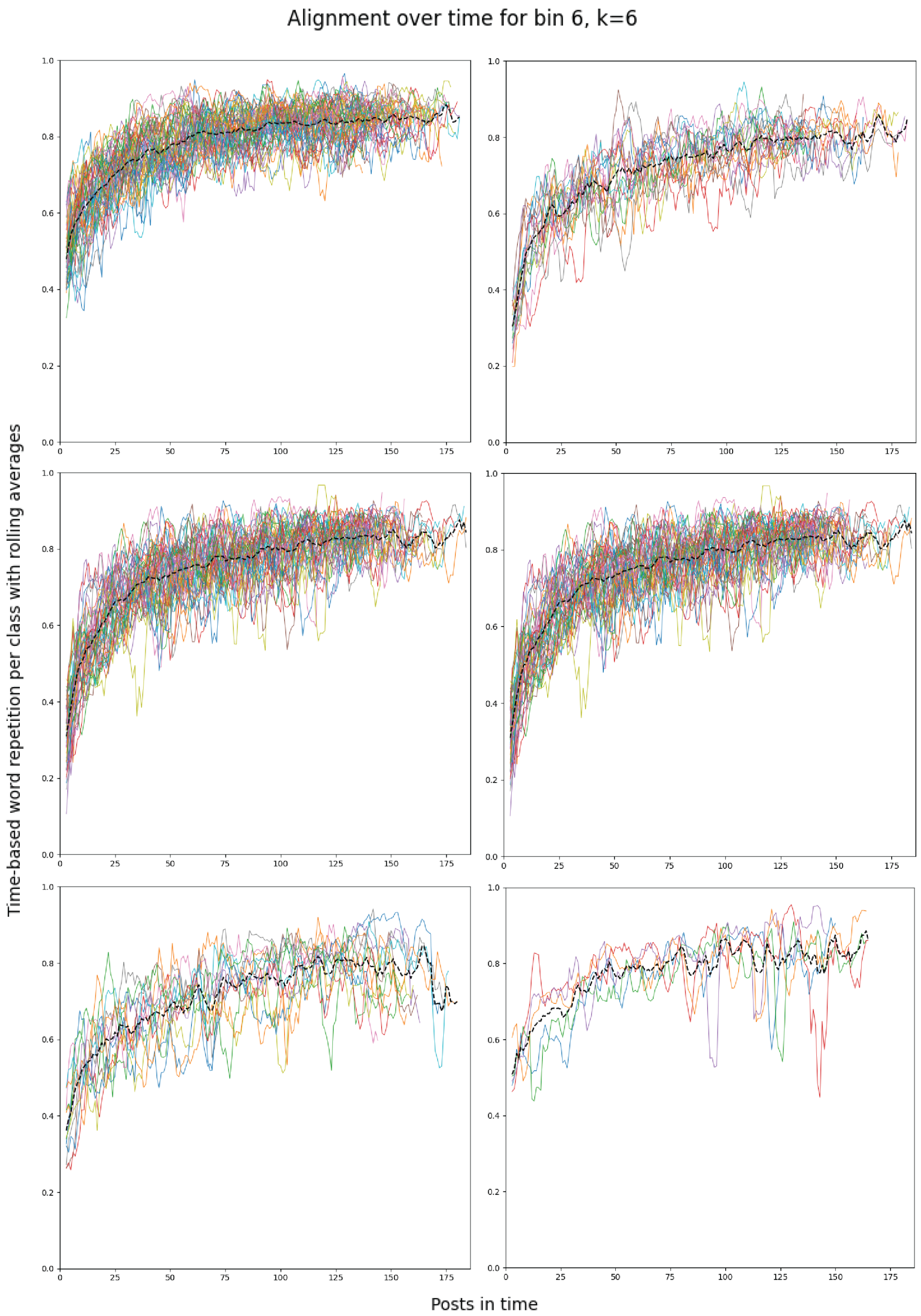


Figure C.11: Clustering results for bin 6, $k = 6$

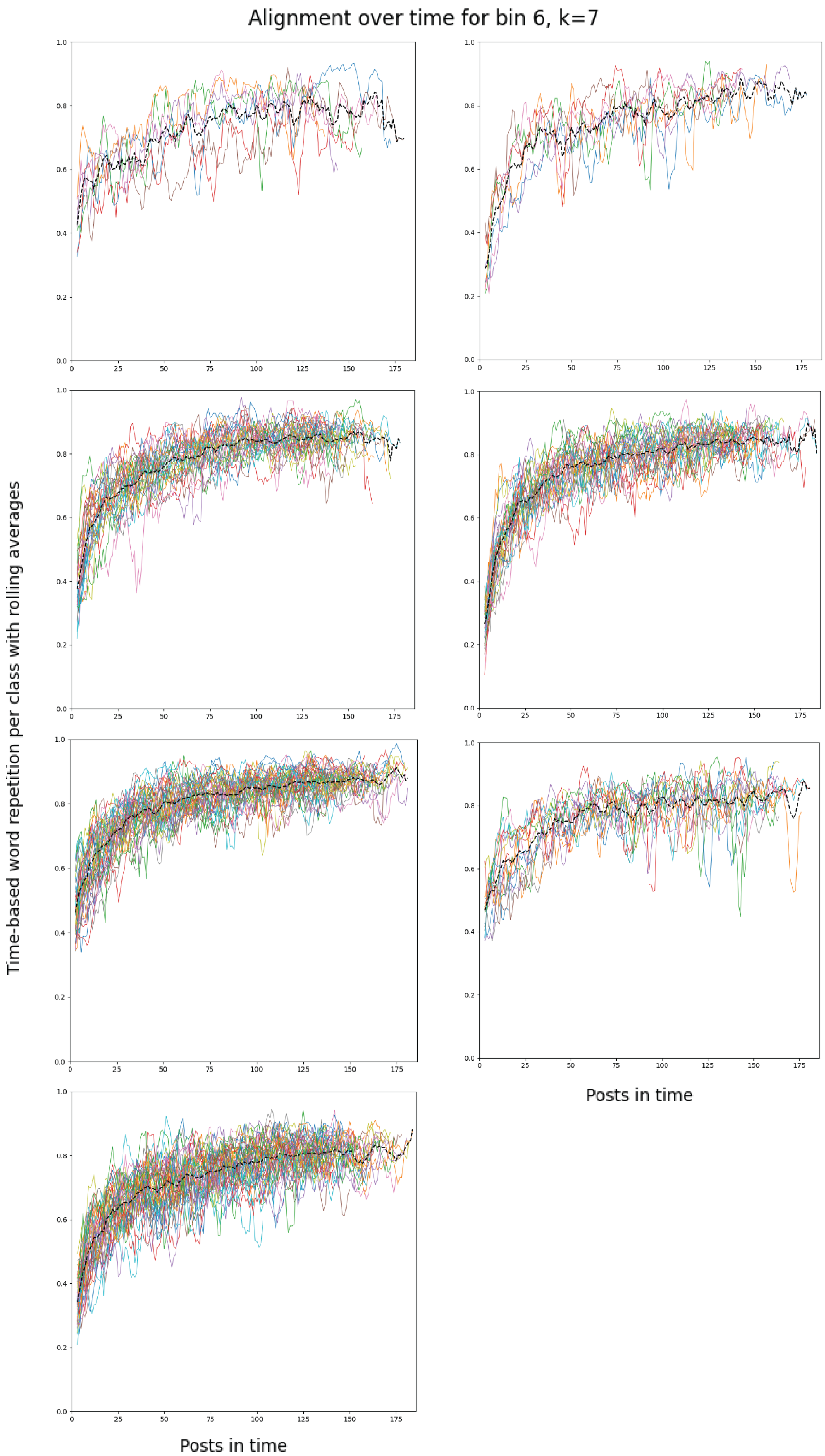


Figure C.12: Clustering results for bin 6, $k = 7$

C.3 Time-based overlap in bin 7

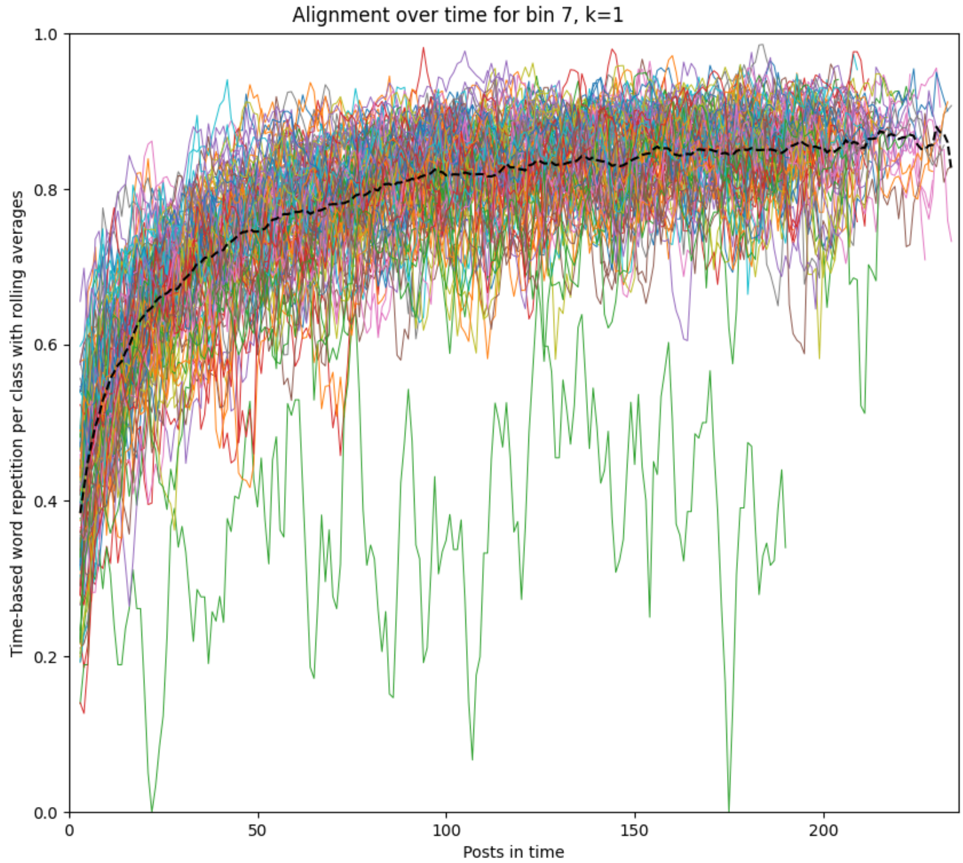


Figure C.13: Rolling average of time-based overlap for bin 7

C.4 Clustering results highlights for bin 8

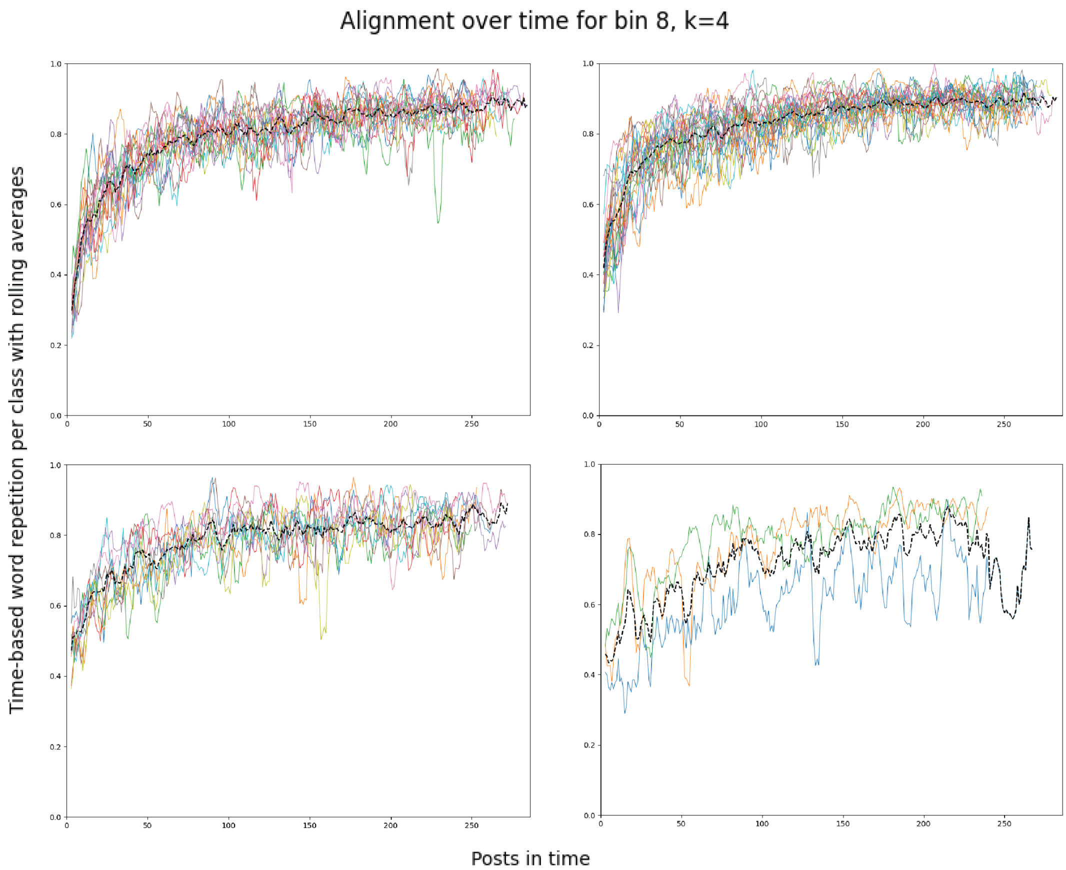


Figure C.14: Clustering results for bin 8, $k = 4$

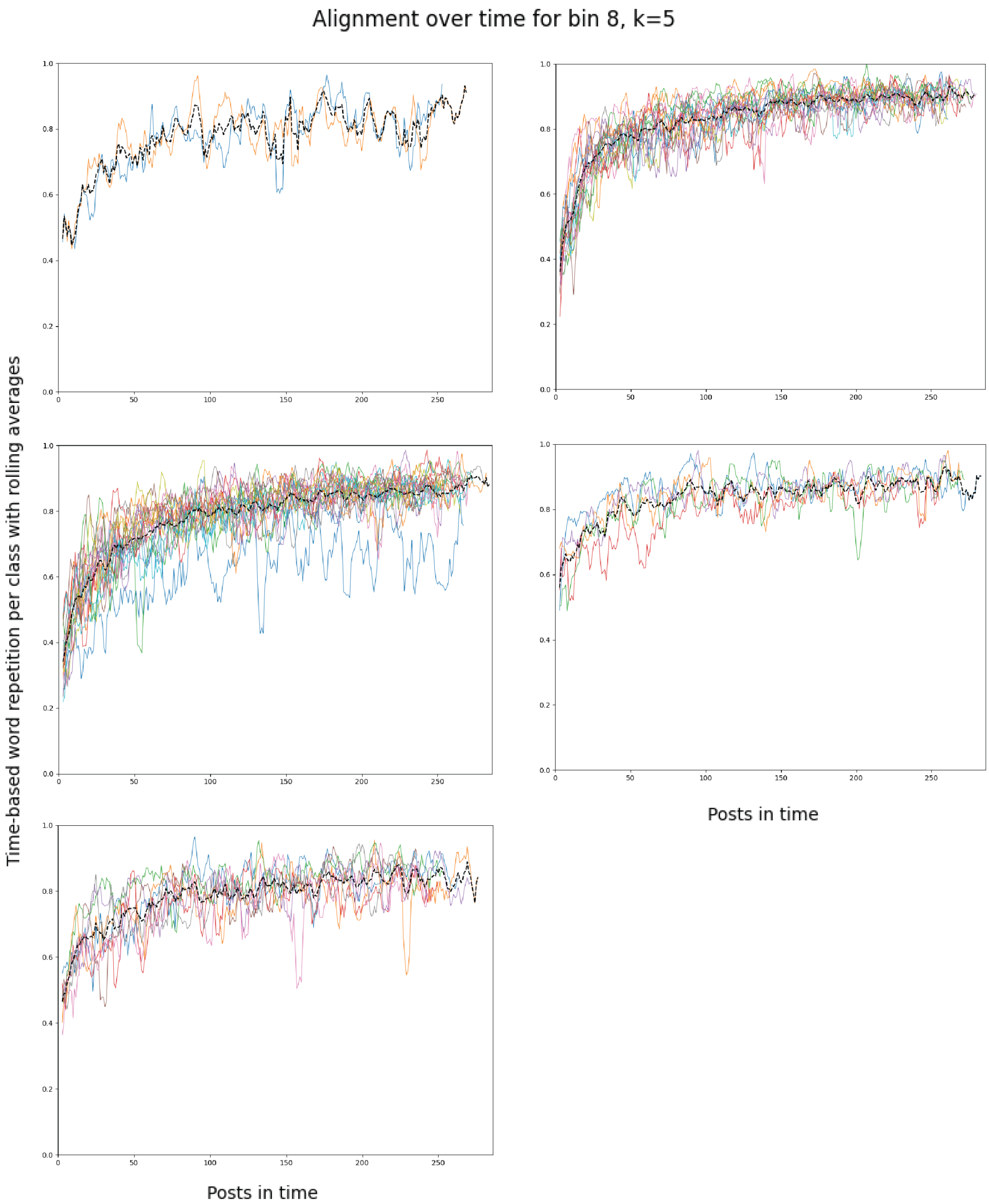


Figure C.15: Clustering results for bin 8, $k = 5$

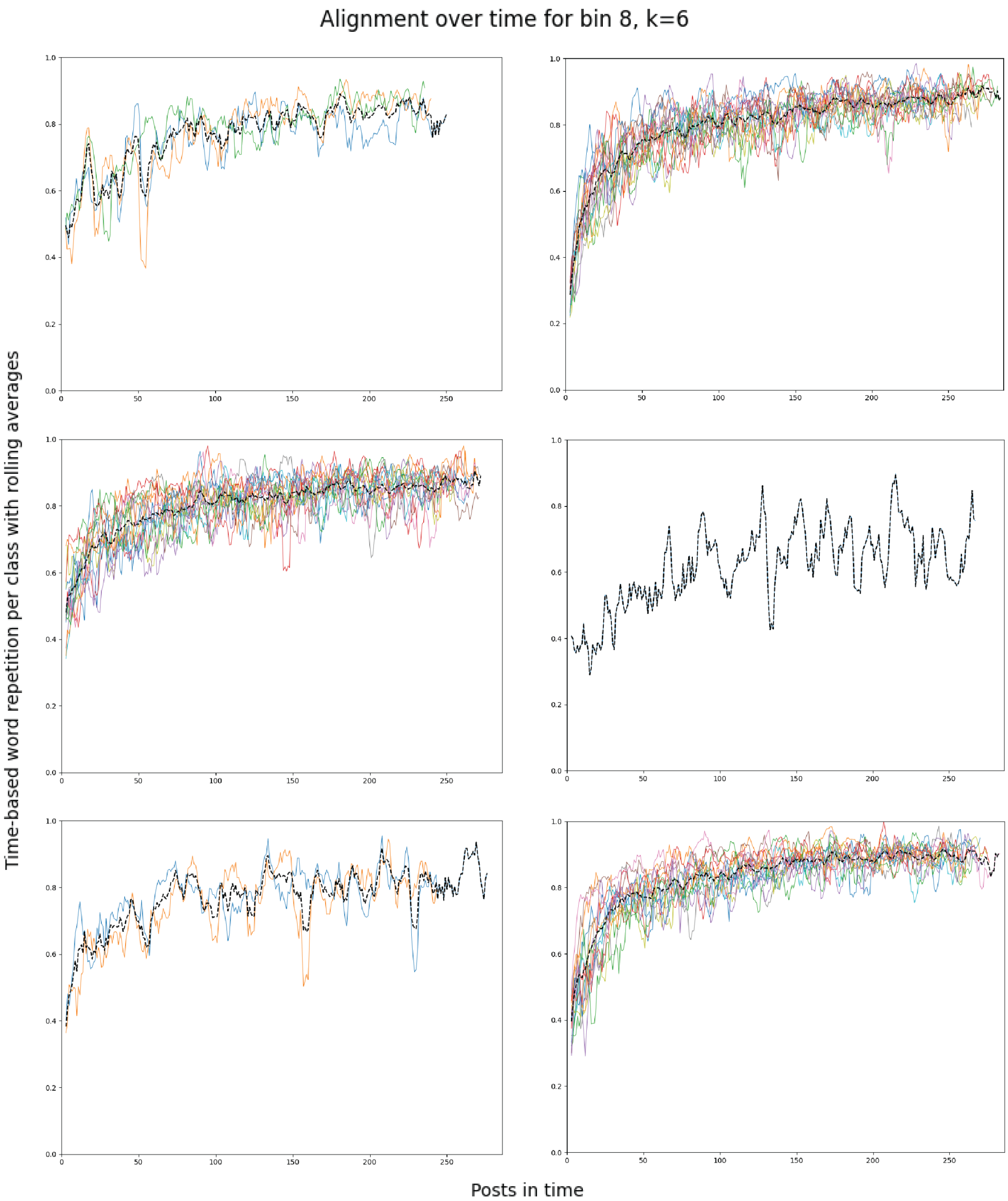


Figure C.16: Clustering results for bin 8, $k = 6$

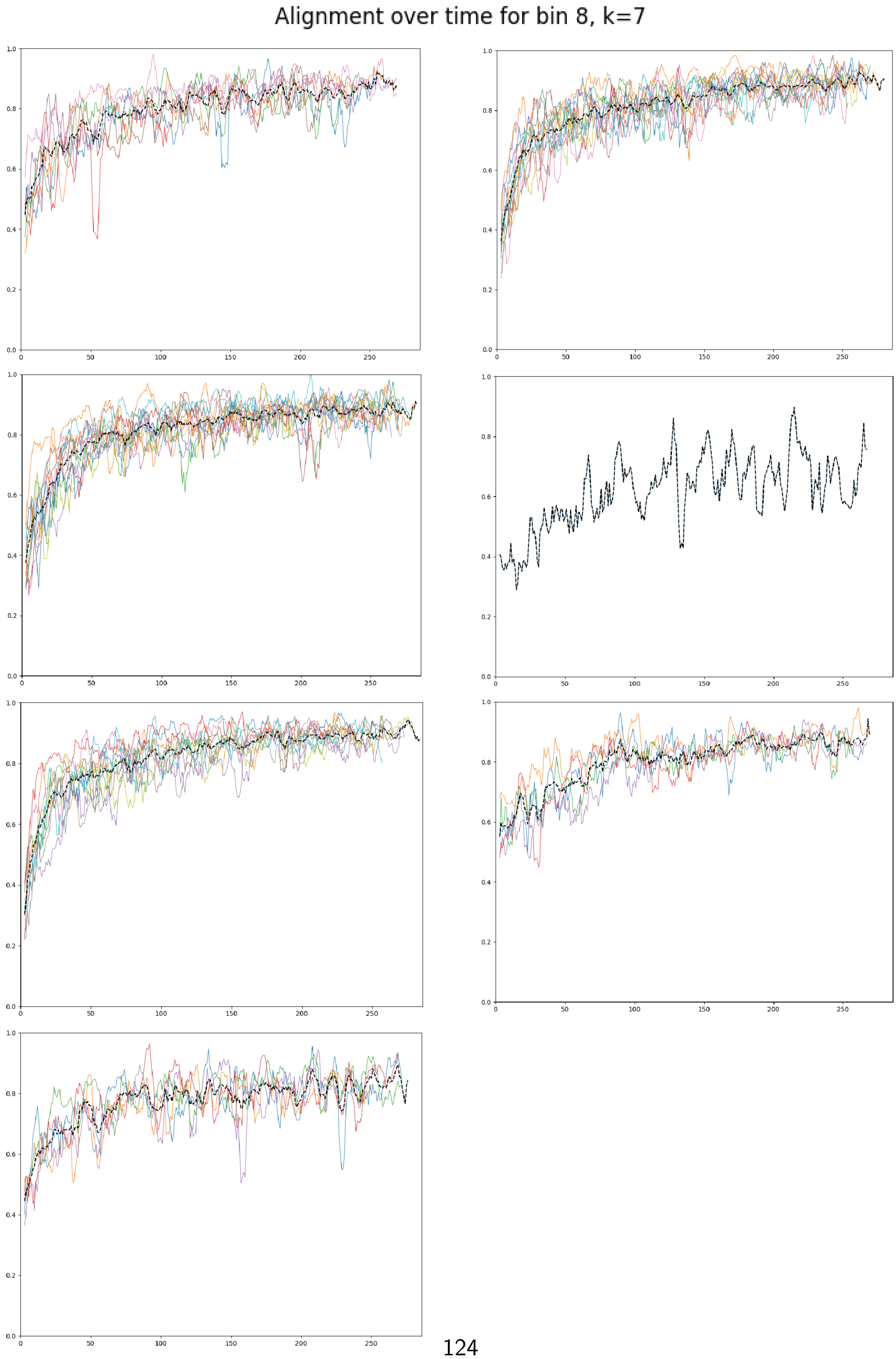


Figure C.17: Clustering results for bin 8, $k = 7$

Appendix D

Additional sentiment results

This appendix contains additional (less insightful) results of the sentiment analysis.

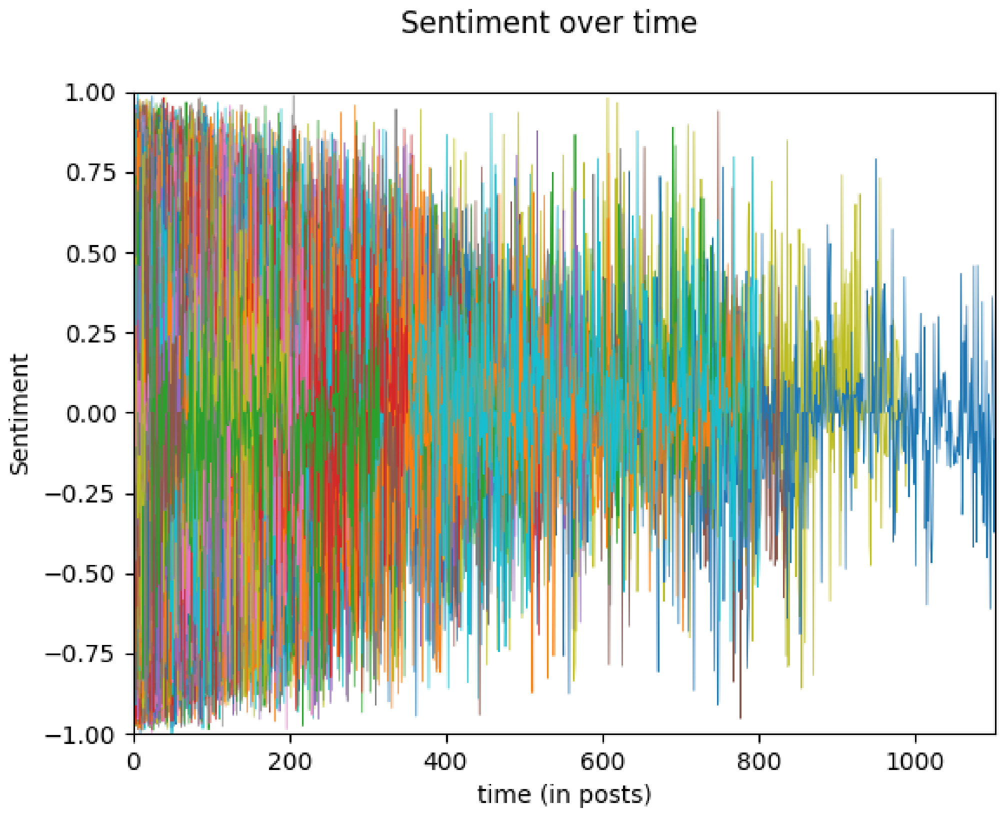


Figure D.1: Sentiment scores changing over posts for all discussions

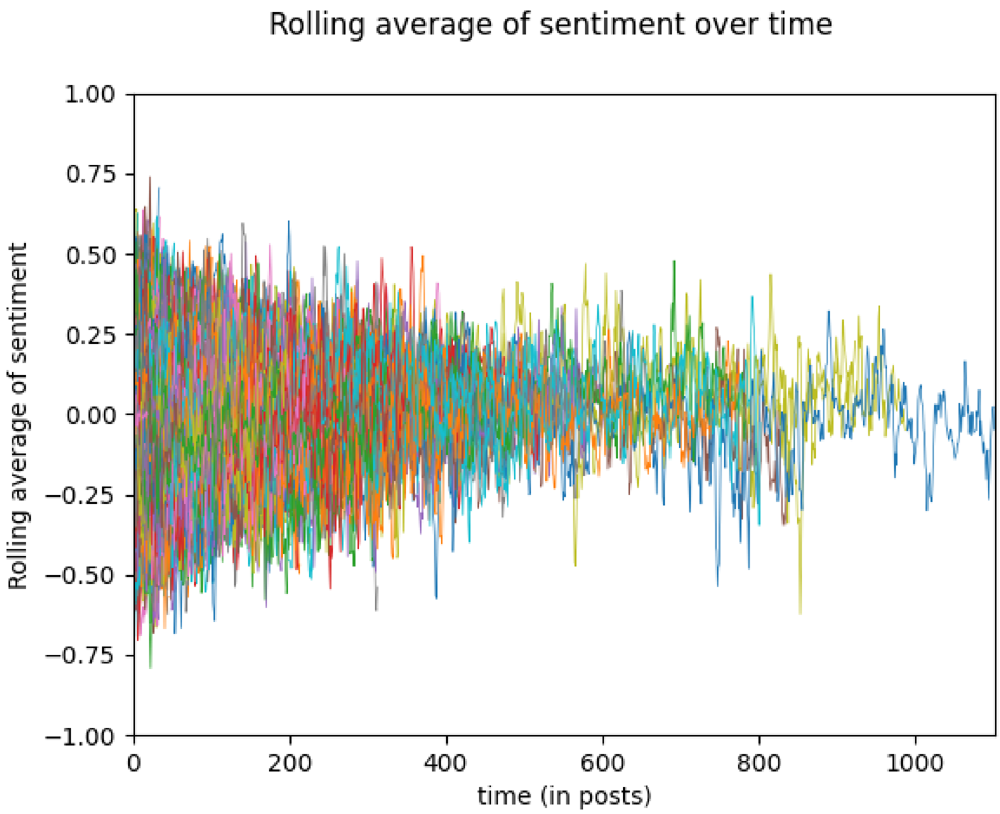


Figure D.2: Rolling averages of sentiment scores changing over posts for all discussions

Appendix E

Clustering finetuning for sentiment

This appendix contains plots that were used to finetune the clustering on sentiment scores.

E.1 Inertia per k for all discussion length bins

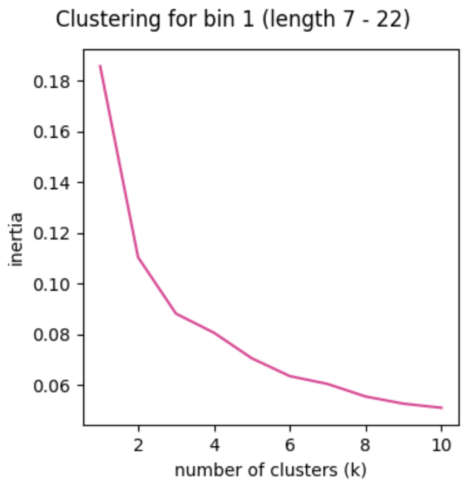


Figure E.1: Inertia per k for the first discussion length bin

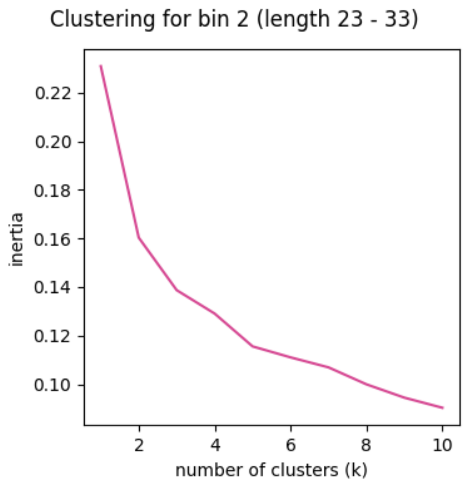


Figure E.2: Inertia per k for the second discussion length bin

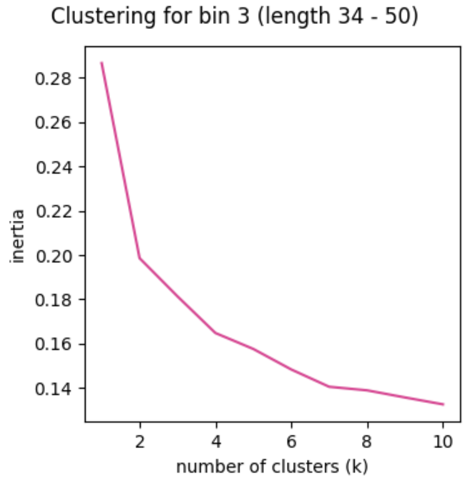


Figure E.3: Inertia per k for the third discussion length bin

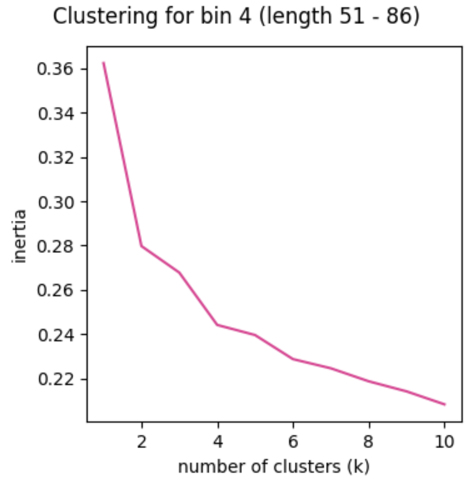


Figure E.4: Inertia per k for the fourth discussion length bin

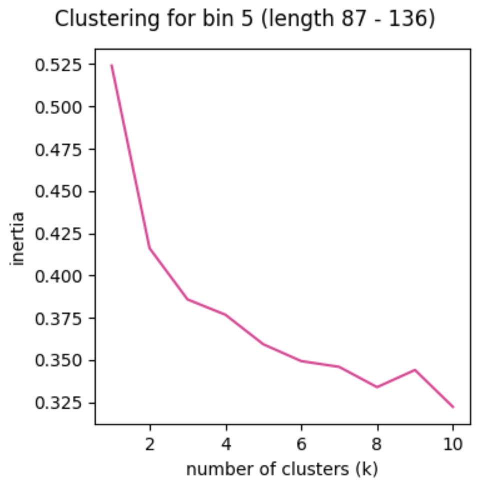


Figure E.5: Inertia per k for the fifth discussion length bin

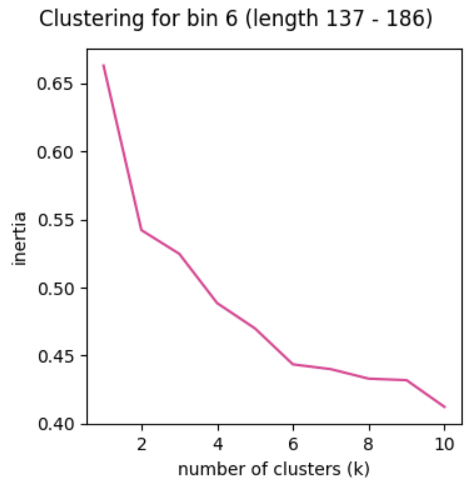


Figure E.6: Inertia per k for the sixth discussion length bin

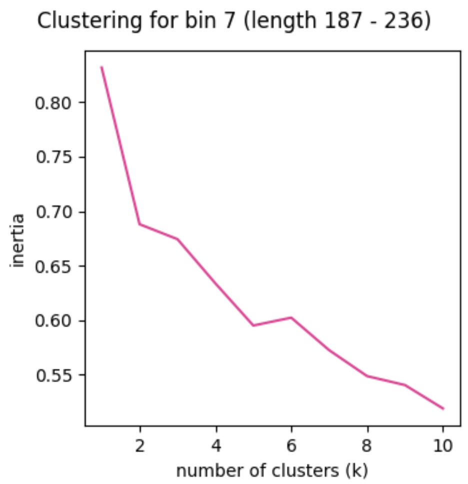


Figure E.7: Inertia per k for the seventh discussion length bin

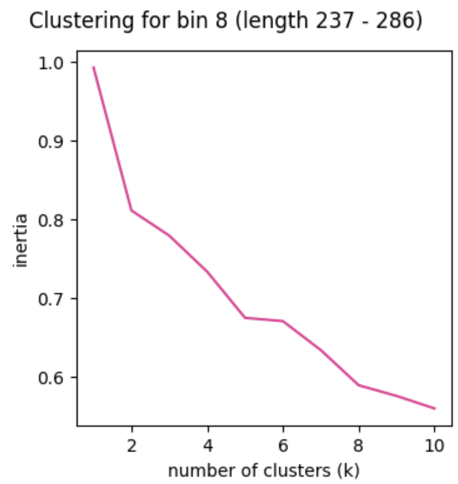


Figure E.8: Inertia per k for the eighth discussion length bin

E.2 Clustering results highlights for bin 5

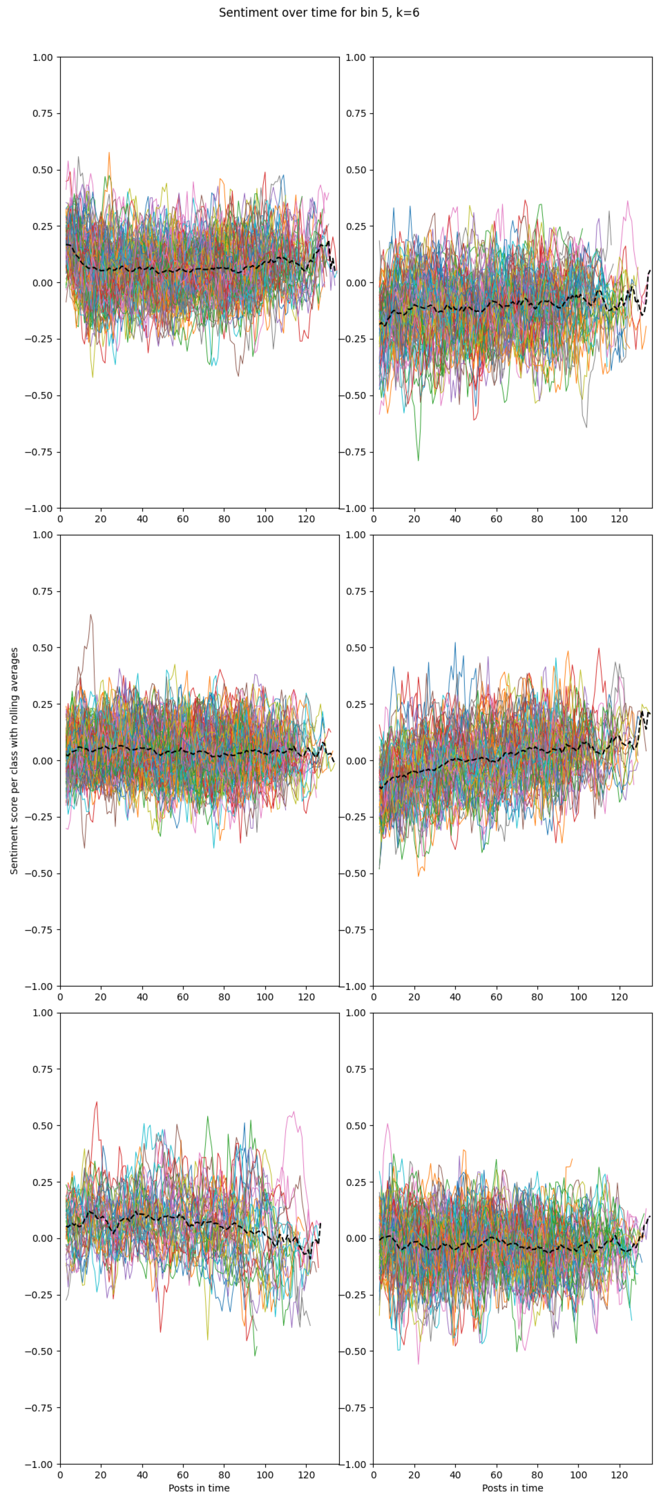


Figure E.9: Clustering results for bin 5, $k = 6$

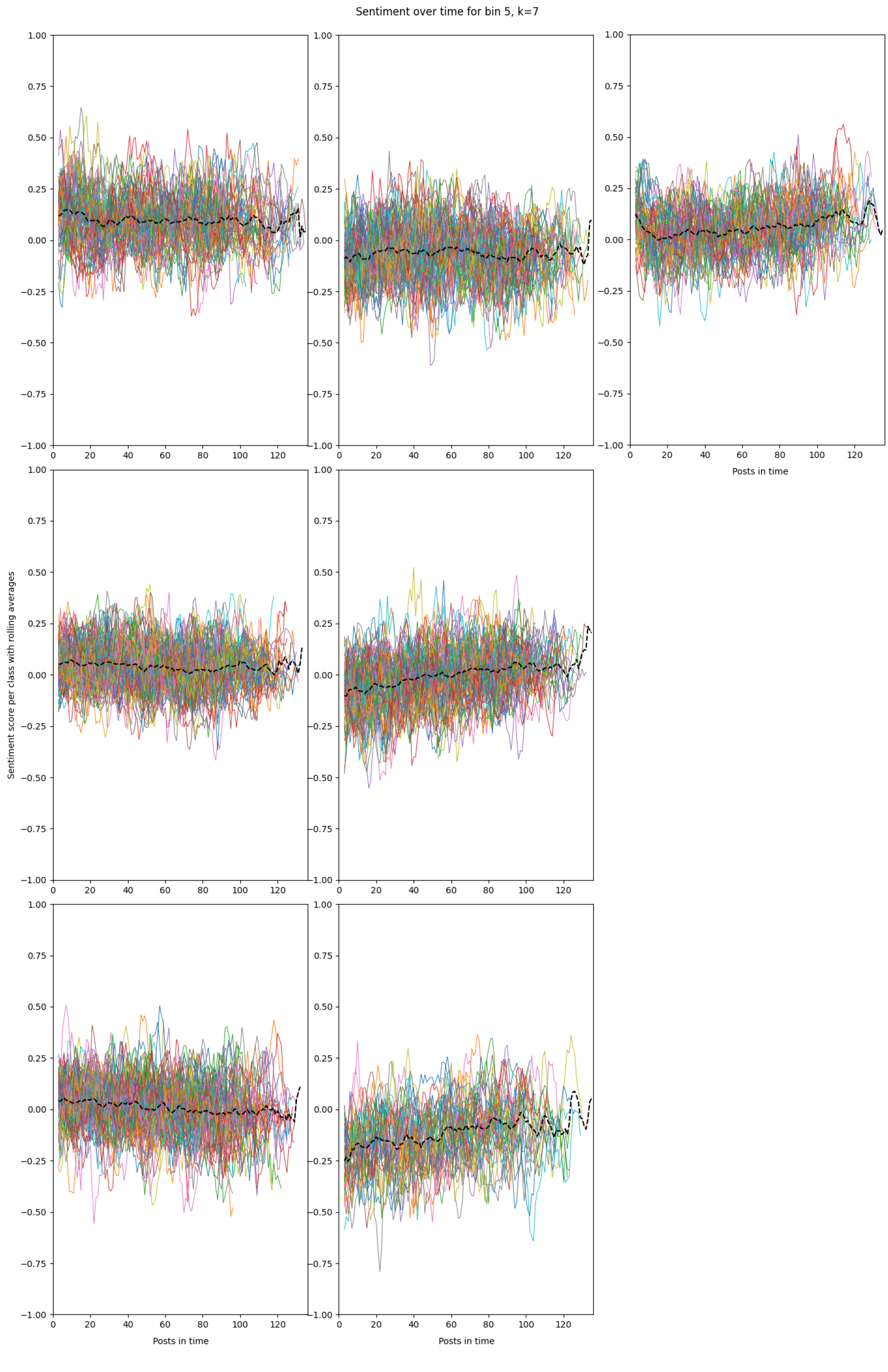


Figure E.10: Clustering results for bin 5, $k = 7$

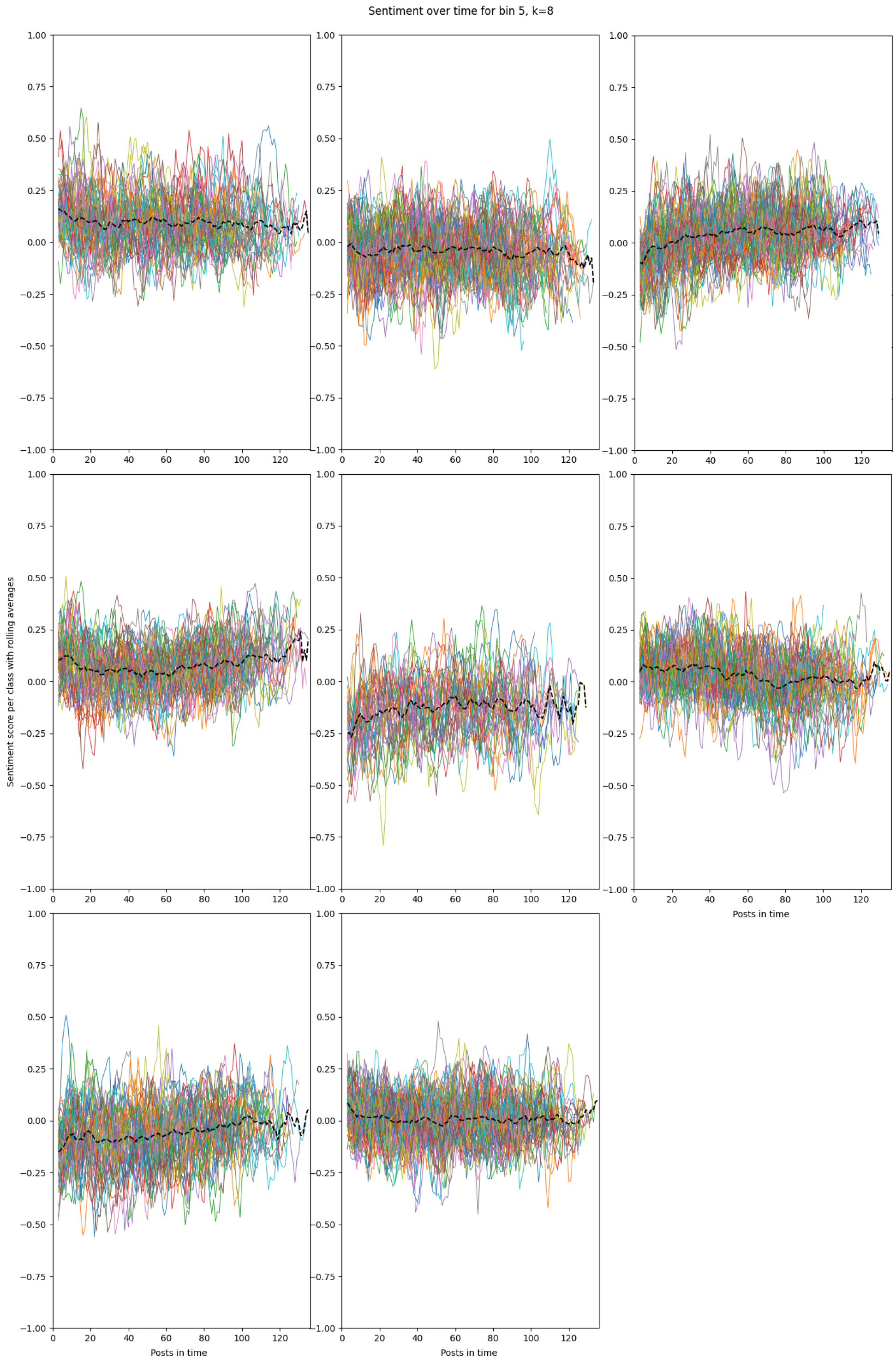


Figure E.11: Clustering results for bin 5, $k = 8$

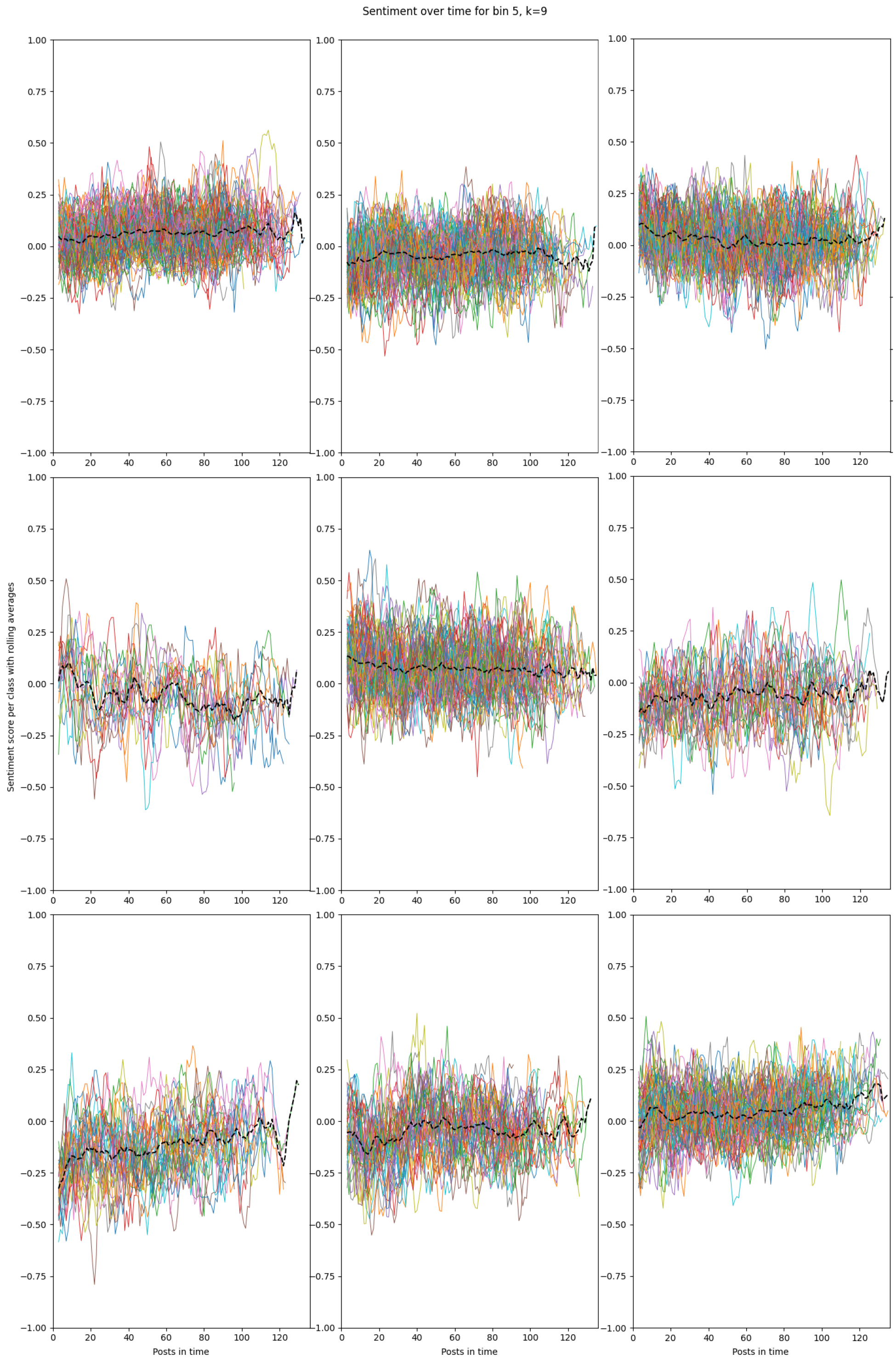


Figure E.12: Clustering results for bin 5, $k = 9$

E.3 Clustering results highlights for bin 7

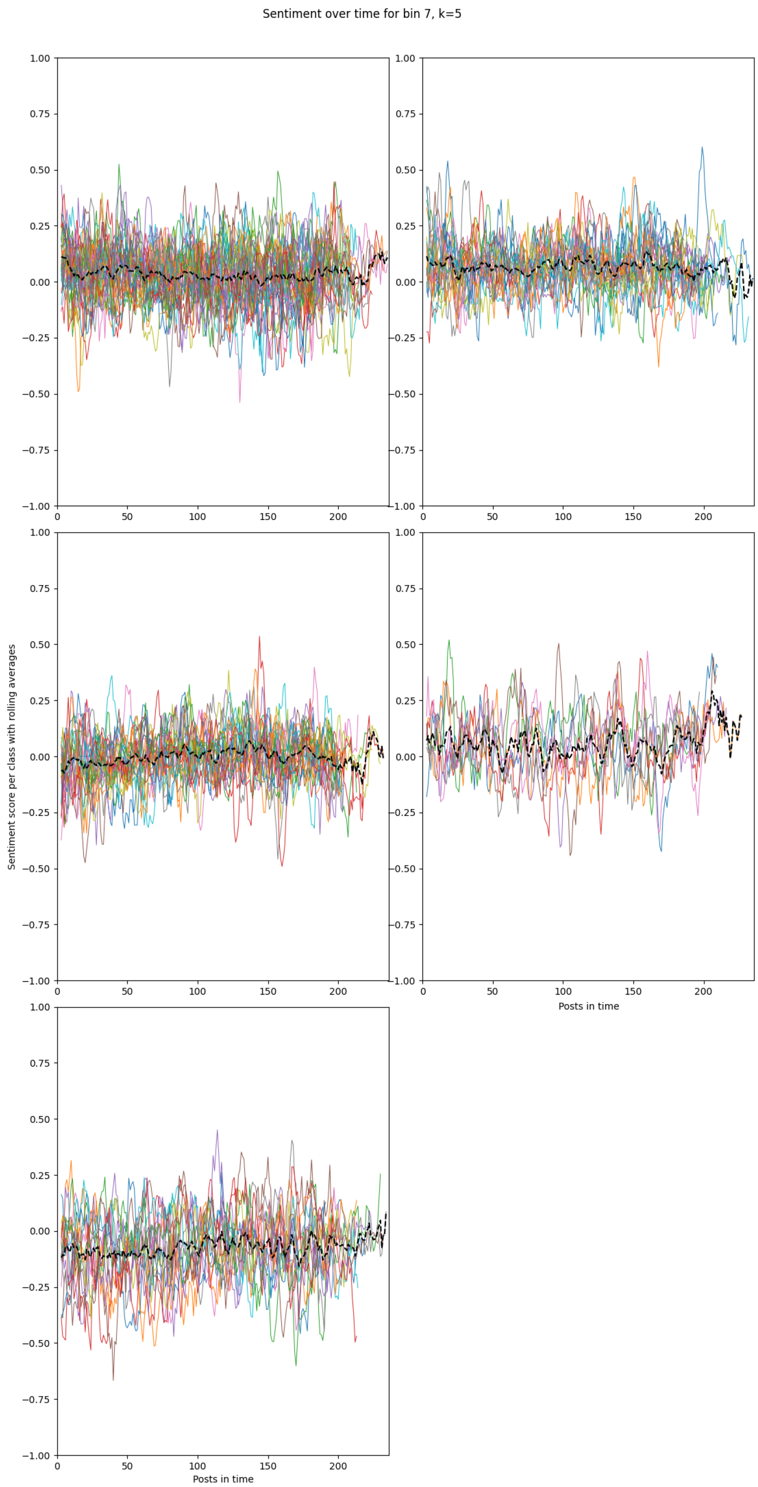


Figure E.13: Clustering results for bin 7, $k = 5$

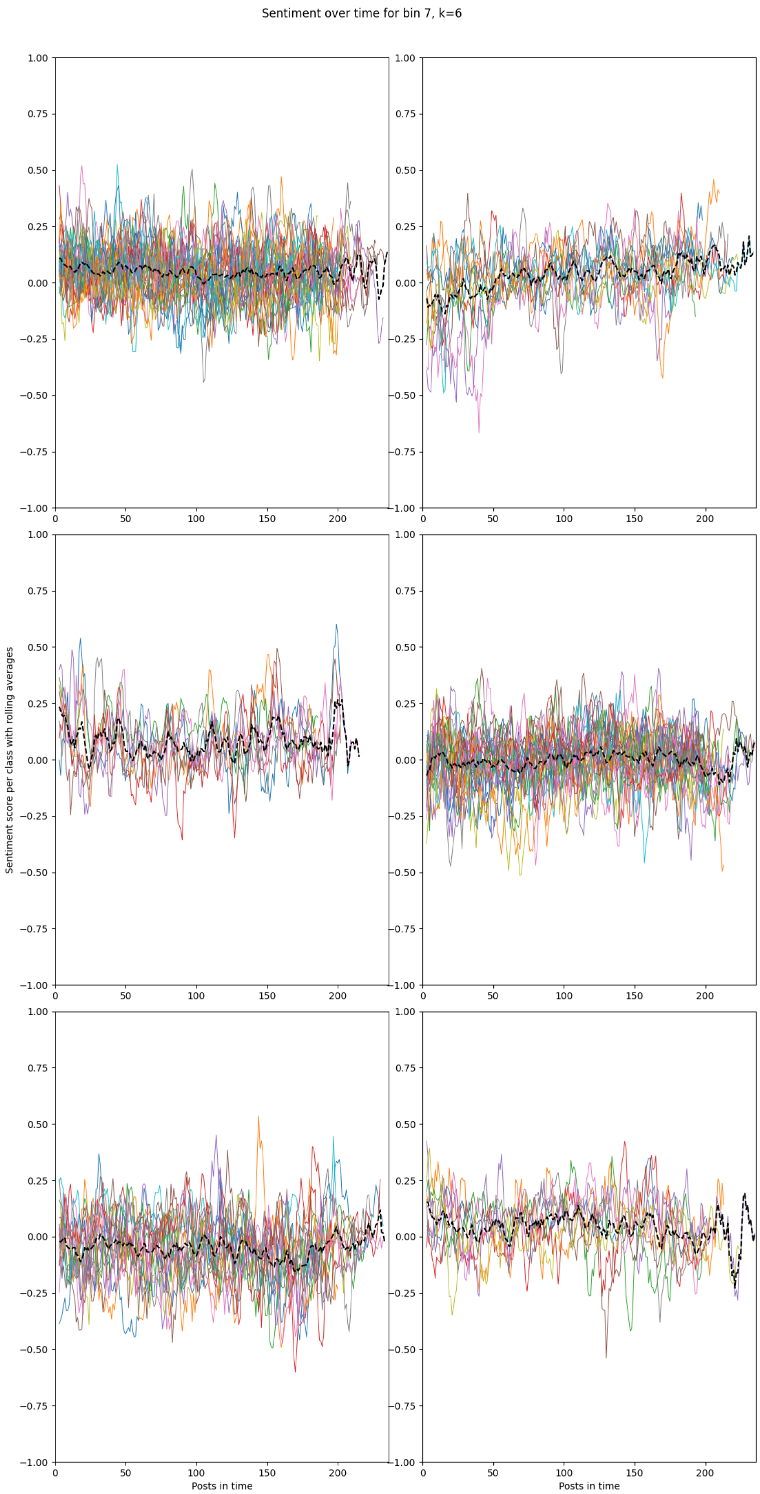


Figure E.14: Clustering results for bin 7, $k = 6$

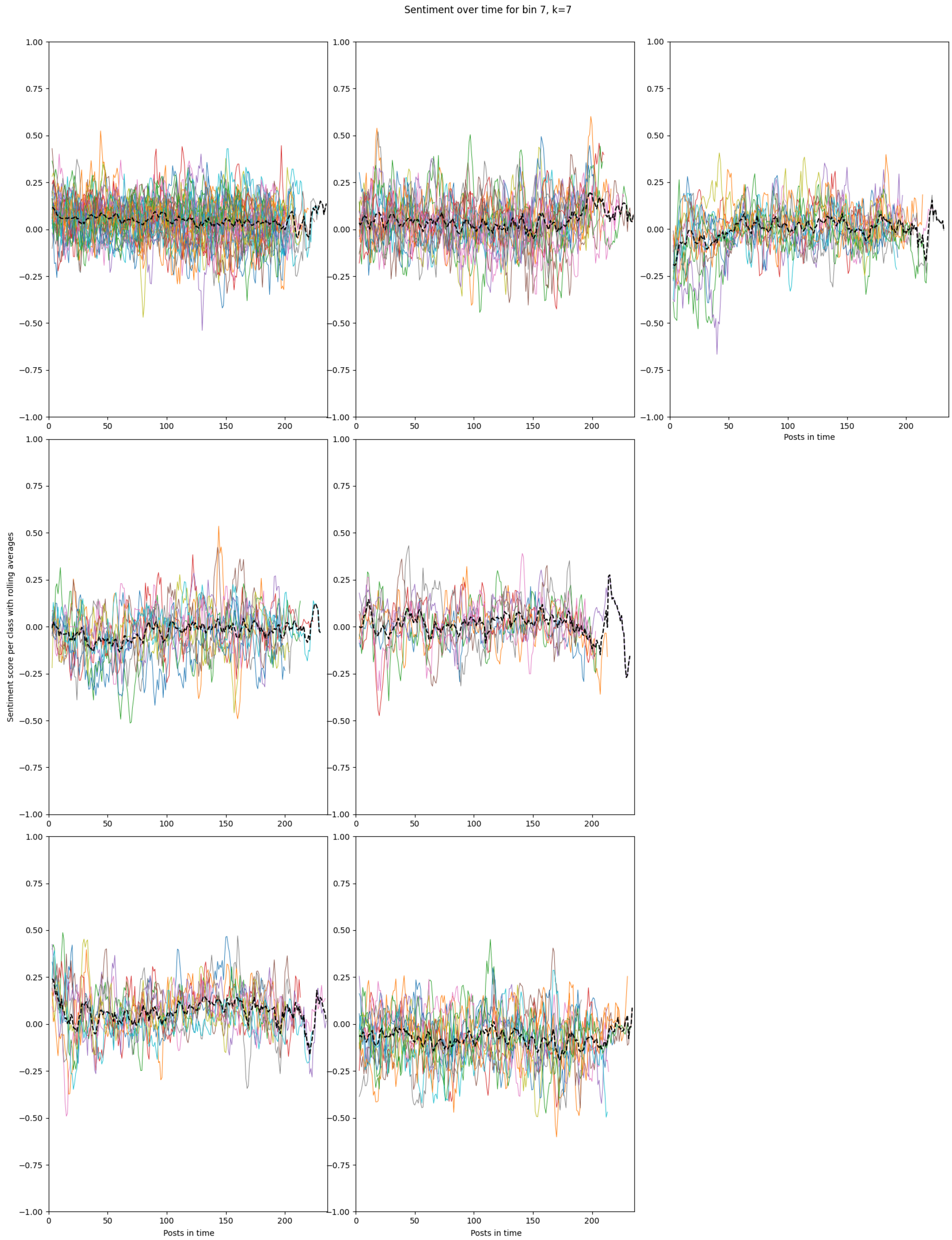


Figure E.15: Clustering results for bin 7, $k = 7$

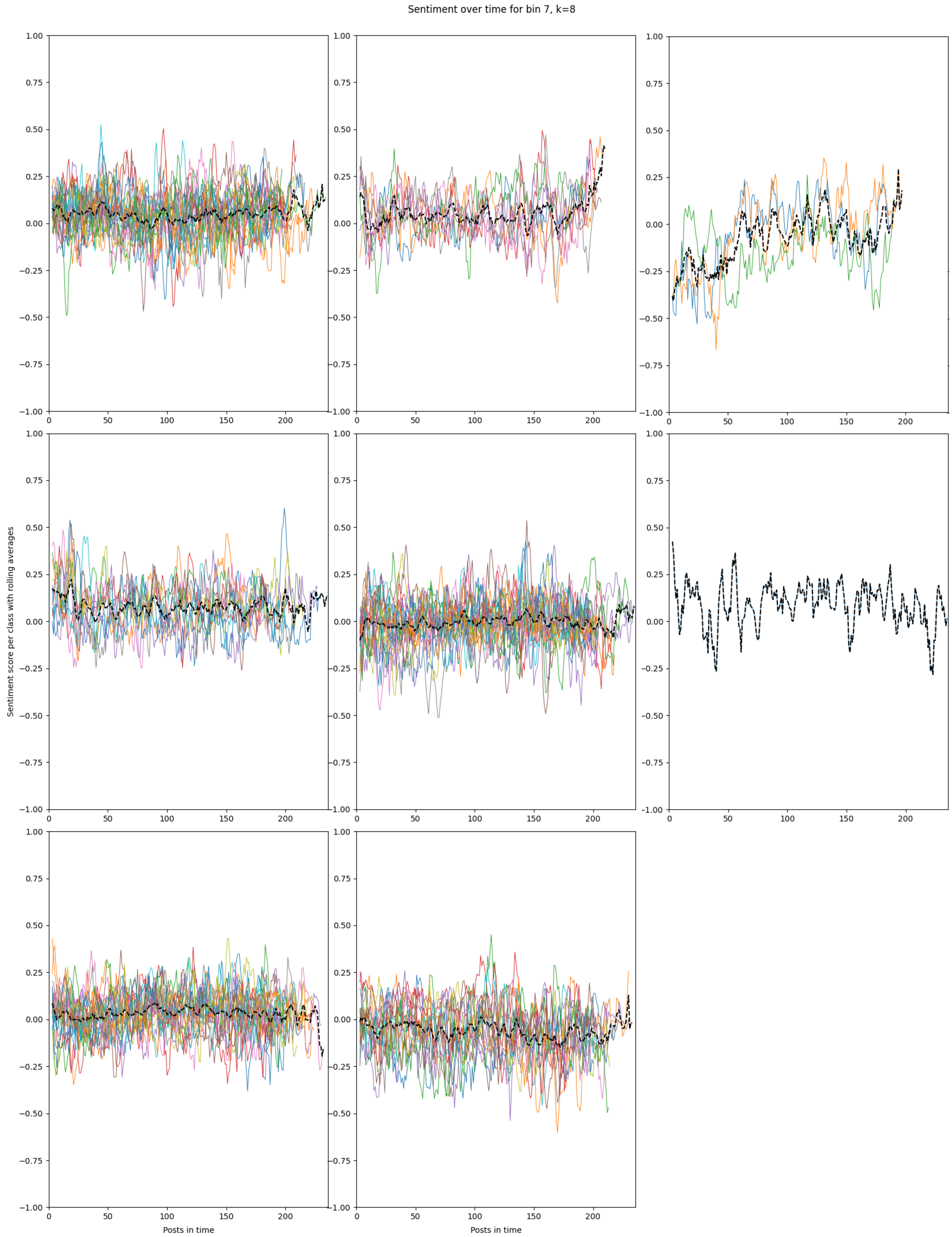


Figure E.16: Clustering results for bin 7, $k = 8$

E.4 Clustering results highlights for bin 8

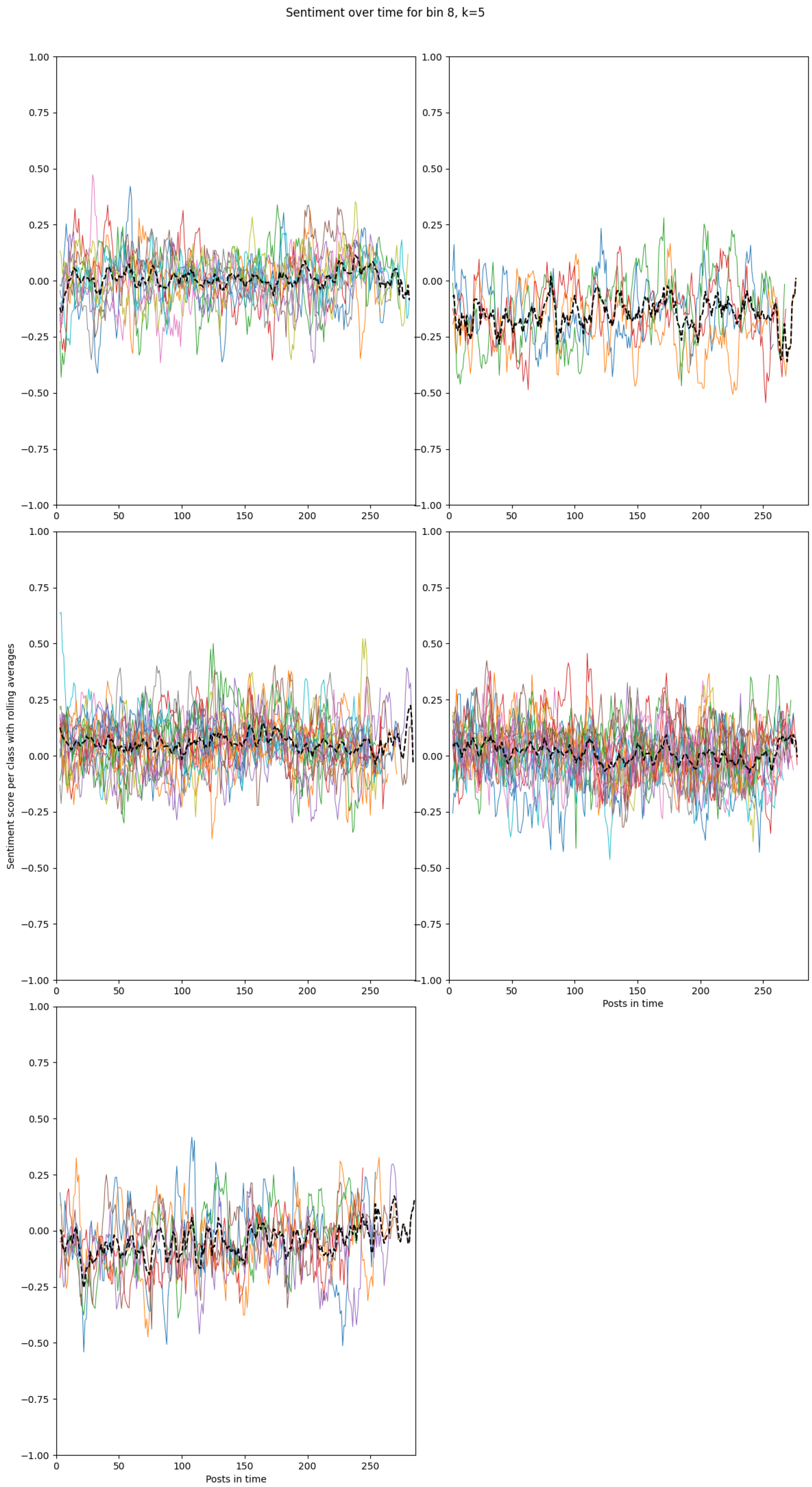


Figure E.17: Clustering results for bin 8, $k = 5$

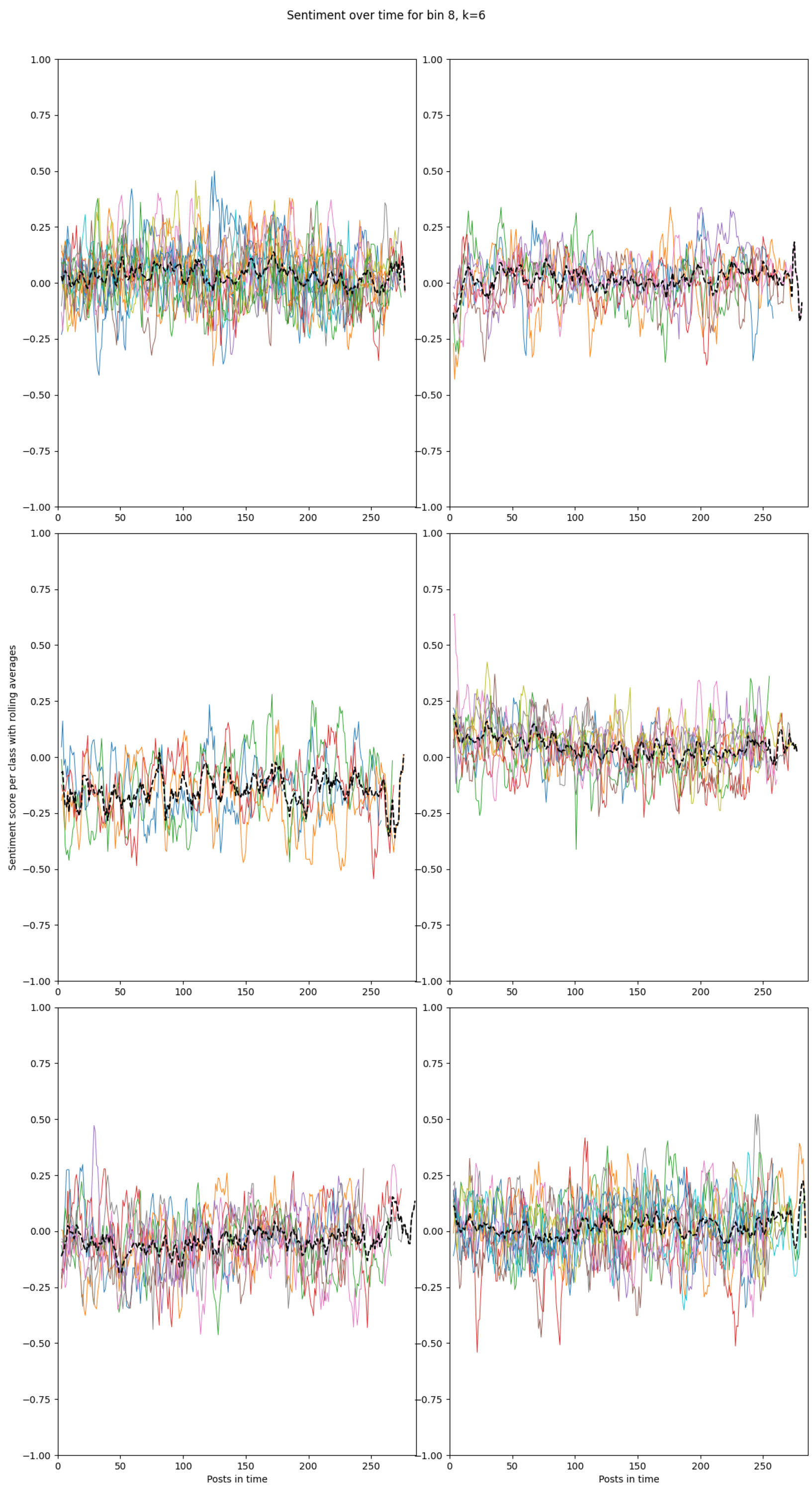


Figure E.18: Clustering results for bin 8, $k = 6$

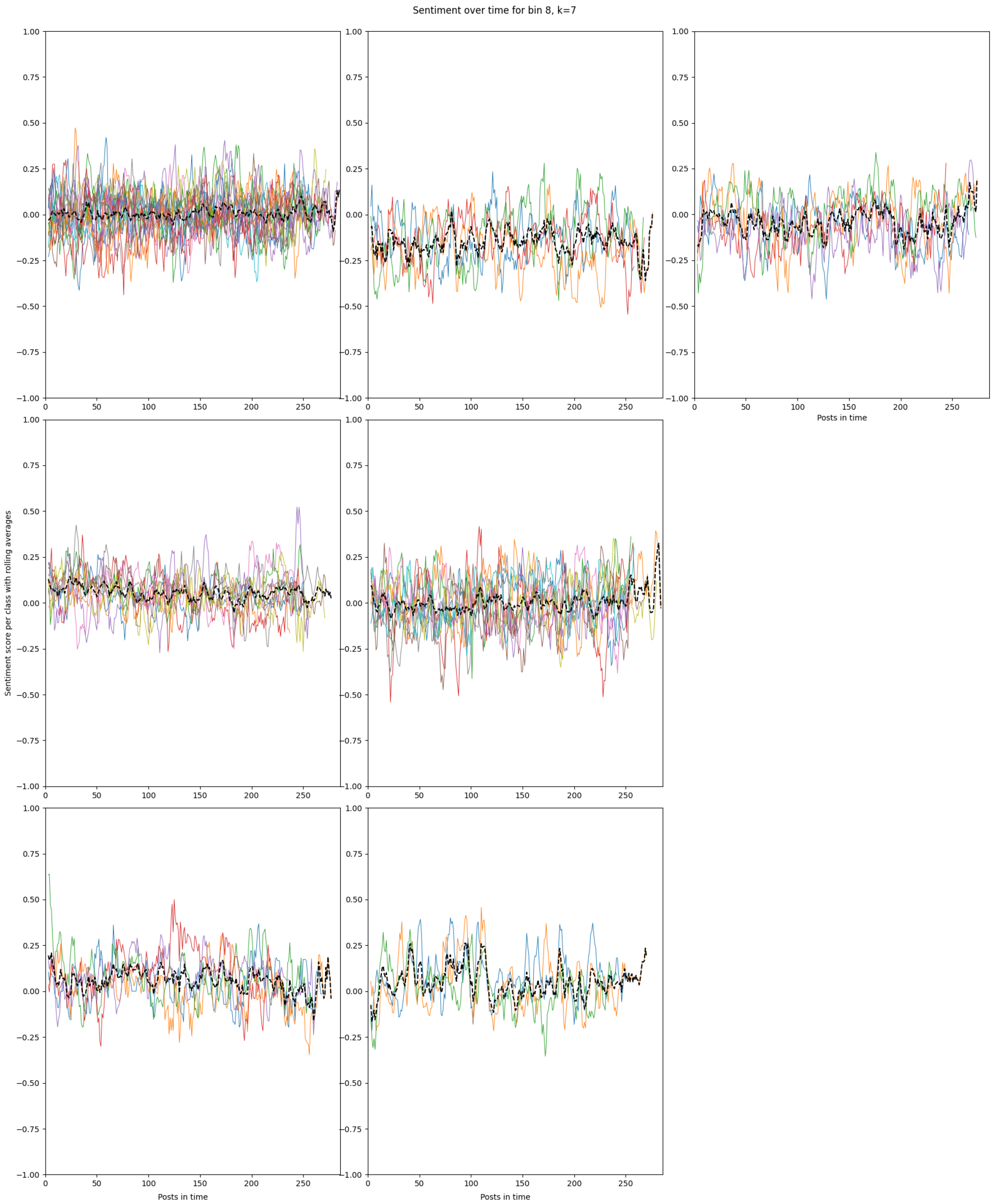


Figure E.19: Clustering results for bin 8, $k = 7$

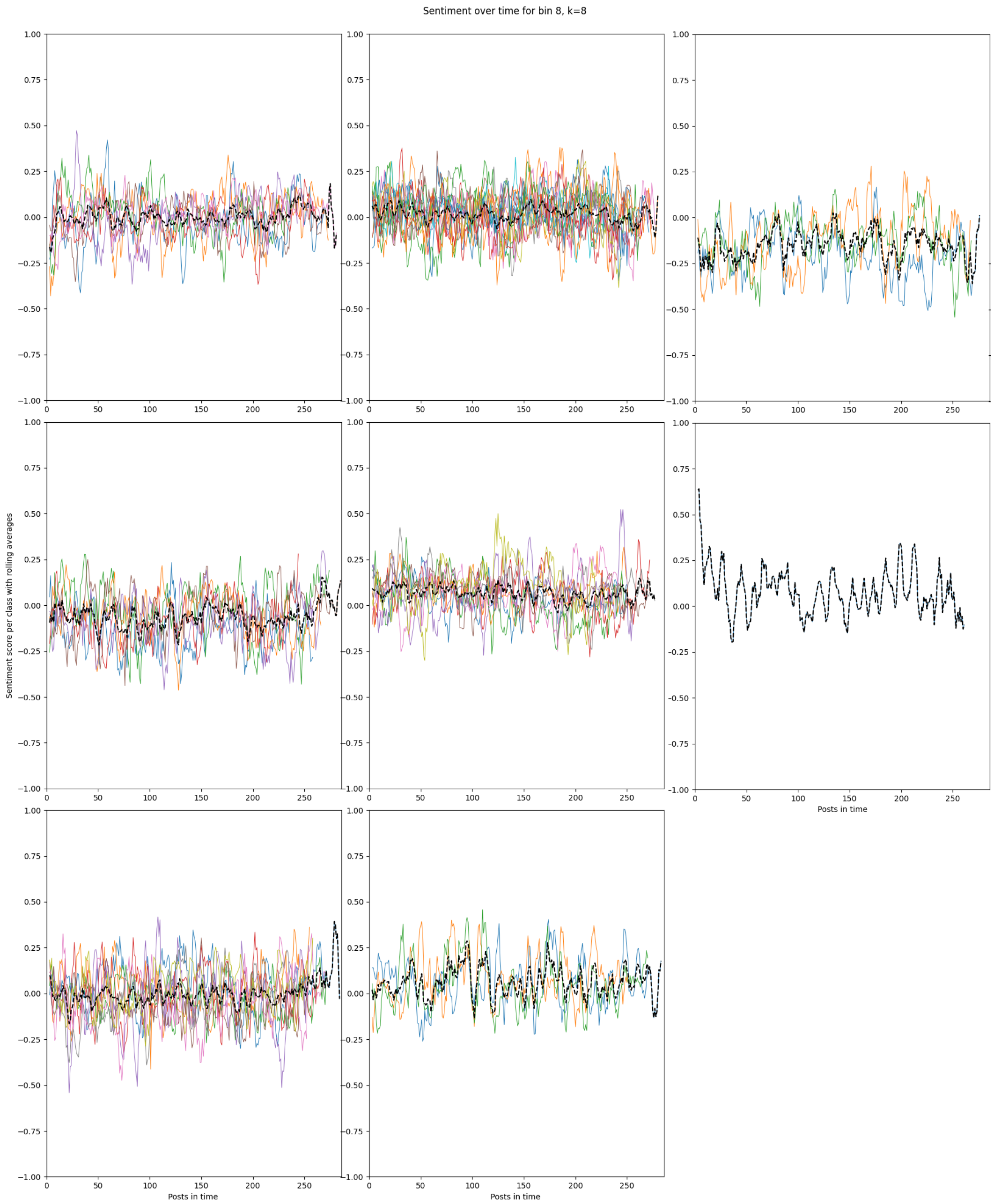


Figure E.20: Clustering results for bin 8, $k = 8$

Appendix F

Contingency tables

F.1 Contingency tables for alignment class vs. sentiment class

This appendix contains the contingency tables with alignment classes against sentiment classes, for each bin of discussion lengths. The contingency tables should be read as follows. Each cell shows how many discussions have that combination of alignment class (group of alignment trends) and sentiment class (group of sentiment trends). The rows and columns are ordered respectively on increasing alignment change and affective shift. A more saturated cell shows a highly populated cell.

		Sentiment classes						**Alignment change**
		4	*6*	*1*	*3*	*2*	*5*	
Alignment classes	*4*	52	20	52	63	67	9	0.145453
	2	50	18	63	40	50	10	0.181052
	1	49	26	55	57	63	18	0.20373
	3	75	36	62	66	65	23	0.297877
Affective shift		-0.12284	-0.08653	-0.05483	0.127225	0.177408	0.228101	

Table F.1: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 1 (7-22 posts)

		Sentiment classes					Alignment change
		5	4	3	1	2	
Alignment classes	2	62	33	35	58	37	0.1804704
	4	35	30	36	61	15	0.2163656
	5	37	18	18	43	12	0.2229463
	6	67	54	26	69	37	0.3299169
	1	35	30	37	56	16	0.3702875
	3	24	15	16	23	14	0.3840732
Affective shift		-0.1039	-0.10302	0.042708	0.115959	0.203079	

Table F.2: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 2 (23-33 posts)

		Sentiment classes							Alignment change
		1	2	6	7	4	3	5	
Alignment classes	2	37	18	39	10	42	23	30	0.2201436
	4	29	14	24	6	39	14	14	0.2265328
	5	21	27	14	8	29	26	22	0.326283
	1	71	42	46	9	62	44	25	0.4169202
	3	56	39	37	15	56	40	30	0.4294556
Affective shift		-0.10933	-0.07861	-0.04148	0.048639	0.064126	0.202685	0.269407	

Table F.3: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 3 (34-50 posts)

		Sentiment classes						Alignment change
		6	2	4	3	5	1	
Alignment classes	2	56	59	18	57	40	24	0.3238632
	1	60	55	8	36	20	24	0.339139
	4	14	15	7	13	7	9	0.3474578
	3	69	62	17	54	31	19	0.4353458
	5	71	74	19	61	35	18	0.5182979
Affective shift		-0.06561	-0.00523	0.159404	0.169169	0.224891	0.25188	

Table F.4: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 4 (51-86 posts)

		Sentiment classes								Alignment change
		5	1	7	3	2	4	8	6	
Alignment classes	4	13	15	18	13	19	8	14	13	0.258296
	3	18	8	12	10	5	12	10	11	0.417103
	2	28	27	43	19	37	12	17	12	0.428433
	1	28	8	24	21	23	21	19	18	0.548944
Affective shift		-0.15862	-0.13856	0.055734	0.069402	0.104009	0.113944	0.133831	0.233061	

Table F.5: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 5 (87-136 posts)

		Sentiment classes						Alignment change
		4	6	3	5	2	1	
Alignment classes	3	1	0	1	2	4	1	0.282995
	5	3	0	6	9	5	1	0.315733
	4	3	7	6	12	15	7	0.473639
	2	14	2	6	9	18	4	0.505759
	1	11	2	4	10	7	4	0.600344
Affective shift		-0.13589	-0.08988	0.004629	0.078875	0.144985	0.297916	

Table F.6: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 6 (137-186 posts)

		Sentiment classes							Alignment change
		4	6	3	1	2	7	5	
Alignment classes	2	0	0	0	0	1	0	0	0.2004361
	3	3	1	0	0	1	4	0	0.4096909
	1	2	7	2	2	4	4	4	0.4098251
	4	6	11	3	7	4	1	3	0.500651
	5	2	9	1	4	8	6	6	0.5299817
Affective shift		-0.09579	-0.07857	-0.07756	0.037595	0.143919	0.198034	0.228877	

Table F.7: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 7 (187-236 posts)

		Sentiment classes						Alignment change
		2	*1*	*4*	*5*	*3*	*6*	
Alignment classes	*3*	4	5	1	1	3	0	0.3241663
	4	1	0	0	0	1	1	0.4518053
	2	1	6	5	3	9	3	0.5336634
	1	0	1	2	0	3	2	0.6014976
Affective shift		-0.24329	-0.04838	0.030908	0.083145	0.165642	0.29259	

Table F.8: Contingency table showing the interrelation between sentiment classes and alignment classes, bin 8 (237-286 posts)

F.2 Contingency tables for sparse bins for alignment classes vs topic

		Topics				Alignment change
		abortion	*evolution*	*gay marriage*	*gun control*	
Alignment classes	*3*	1	0	1	0	0.409691
	1	2	3	3	6	0.409825
	4	2	4	4	0	0.500651
	5	5	6	2	1	0.529982

Table F.9: Contingency table showing the interrelation between alignment classes and topic, bin 7 (187-236 posts)

		Topics			Alignment change
		abortion	*evolution*	*gun control*	
Alignment classes	*3*	2	0	1	0.3241663
	4	0	0	1	0.4518053
	2	4	8	3	0.5336634
	1	2	0	0	0.6014976

Table F.10: Contingency table showing the interrelation between alignment classes and topic, bin 8 (237-286 posts)

	Bin 7	*Bin 8*
Cramér's V for all data	0.4212	0.5921
Cramér's V without sparse data	0.3405	0.4848

Table F.11: Cramér's V for the last two bins for alignment classes and topics

F.3 Original contingency tables for alignment classes vs topic

		Topics								
		abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gun control	socialized medicine
Alignment classes	1	12	0	2	0	14	1	2	30	1
	2	13	1	2	1	35	1	8	17	0
	3	14	4	0	2	19	3	9	29	0
	4	7	3	0	0	19	2	4	21	0

Table F.12: Sparse contingency table showing the interrelation between alignment classes and topic, bin 1 (7-22 posts)

		Topics											
		abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gun control	marijuana legalization	minimum wage: pro or con	obamacare	women in the military
	1	10	0	0	0	17	3	10	14	0	1	1	0
	2	17	1	0	0	16	0	2	19	0	0	0	0
	3	2	1	1	1	4	1	3	10	0	0	0	1
	4	5	3	0	0	9	0	3	16	0	1	0	0
Alignment	5	5	1	0	0	20	4	7	7	0	0	0	0
classes	6	10	5	3	1	14	3	11	30	1	0	0	1

Table F.13: Sparse contingency table showing the interrelation between alignment classes and topic, bin 2 (23-33 posts)

		abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gun control	immigration	minimum wage: pro or con	obamacare
Alignment classes	1	16	2	1	1	42	5	10	18	1	1	0
	2	9	3	0	2	12	3	5	21	0	0	0
	3	12	3	5	2	18	3	8	20	1	0	1
	4	15	0	1	0	16	6	4	10	0	1	1
	5	4	0	1	0	5	0	2	14	0	0	0

Table F.14: Sparse contingency table showing the interrelation between alignment classes and topic, bin 3 (34-50 posts)

		Topics											
		abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gun control	legalized prostitution	marijuana legalization	obamacare	women in the military
Alignment classes	1	15	1	1	1	33	6	13	13	0	1	1	0
	2	19	0	0	1	20	7	10	25	1	0	0	0
	3	15	2	2	2	19	2	7	32	0	1	1	0
	4	7	0	0	0	3	0	1	5	0	0	0	0
	5	26	0	2	0	27	4	10	32	0	0	1	2

Table F.15: Sparse contingency table showing the interrelation between alignment classes and topic, bin 4 (51-86 posts)

		Topics														
		School Uniforms	abortion	climate change	communism vs capitalism	death penalty	evolution	existence of God	gay marriage	gays in the military	gun control	immigration	legalized prostitution	marijuana legalization	obamacare	women in the military
Alignment classes	1	1	18	1	0	1	15	0	7	0	17	1	1	1	0	1
	2	0	24	1	0	2	28	10	6	0	15	0	0	1	0	0
	3	0	5	1	1	2	3	0	3	0	6	0	0	0	1	0
	4	0	13	0	0	2	7	3	5	1	11	0	0	0	1	0

Table F.16: Sparse contingency table showing the interrelation between alignment classes and topic, bin 5 (87-136 posts)

		Topics							
		School Uniforms	abortion	climate change	death penalty	evolution	existence of God	gay marriage	gun control
Alignment classes	1	0	3	0	0	1	0	2	1
	2	0	7	1	0	10	0	6	4
	3	0	1	0	0	0	1	0	1
	4	1	4	1	1	4	1	1	10
	5	0	5	0	0	1	1	0	1

Table F.17: Sparse contingency table showing the interrelation between alignment classes and topic, bin 6 (137-186 posts)

		Topics						
		abortion	evolution	existence of God	gay marriage	gun control	legalization	vegetarianism
Alignment classes	1	2	3	0	3	6	1	0
	2	1	0	0	1	0	0	0
	3	2	4	2	4	0	0	0
	4	5	6	0	2	1	0	1

Table F.18: Sparse contingency table showing the interrelation between alignment classes and topic, bin 7 (187-236 posts)

		Topics						
		abortion	climate change	death penalty	evolution	existence of God	gay marriage	gun control
Alignment classes	1	2	1	0	0	0	0	0
	2	4	0	1	8	0	1	3
	3	2	0	0	0	2	0	1
	4	0	0	0	0	0	0	1

Table F.19: Sparse contingency table showing the interrelation between alignment classes and topic, bin 8 (237-286 posts)

F.4 Contingency tables for sparse bins for sentiment classes vs topic

		Topic				*Affective shift*
		abortion	*evolution*	*gay marriage*	*gun control*	
Sentiment classes	*4*	1	2	1	1	-0.09579
	6	3	7	3	0	-0.07857
	3	0	1	0	0	-0.07756
	1	0	0	2	0	0.037595
	2	3	0	2	2	0.143919
	7	2	2	2	0	0.198034
	5	1	1	0	4	0.228877

Table F.20: Contingency table showing the interrelation between sentiment classes and topic, bin 7 (187-236 posts)

		Topics		Affective shift
		abortion	*evolution*	
Sentiment classes	*2*	0	1	-0.24329
	1	2	5	-0.04838
	4	3	0	0.030908
	3	3	2	0.165642

Table F.21: Contingency table showing the interrelation between sentiment classes and topic, bin 8 (237-286 posts)

	Bin 7	*Bin 8*
Cramér's V for all data	0.3932	0.6839
Cramér's V without sparse data	0.4823	0.5855

Table F.22: Cramér's V for the last two bins for sentiment classes and topics

F.5 Original contingency tables for sentiment classes vs topic

		Topics								
		abortion	*climate change*	*communism vs capitalism*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*	*socialized medicine*
Sentiment classes	1	14	2	2	0	31	0	7	11	0
	2	9	3	1	0	12	1	4	23	0
	3	6	2	0	0	23	2	7	3	0
	4	10	1	1	0	20	4	4	10	1
	5	6	0	0	1	1	0	0	14	0
	6	1	0	0	2	0	0	1	36	0

Table F.23: Sparse contingency table showing the interrelation between sentiment classes and topic, bin 1 (7-22 posts)

		Topics											
		abortion	*climate change*	*communism vs capitalism*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*	*marijuana legalization*	*minimum wage: pro or con*	*obamacare*	*women in the military*
Sentiment classes	*1*	13	5	1	1	22	1	16	21	0	0	1	1
	2	7	0	0	1	2	1	2	28	0	0	0	1
	3	5	1	0	0	16	5	5	3	0	1	0	0
	4	7	1	0	0	3	1	5	36	1	0	0	0
	5	17	4	3	0	37	3	8	8	0	1	0	0

Table F.24: Sparse contingency table showing the interrelation between alignment classes and topic, bin 2 (23-33 posts)

		Topics										
		abortion	*climate change*	*communism vs capitalism*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*	*immigration*	*minimum wage: pro or con*	*obamacare*
Sentiment class	*1*	8	3	1	0	30	7	6	6	0	1	0
	2	10	0	0	2	4	0	2	18	0	0	0
	3	13	0	2	0	6	2	3	12	1	0	0
	4	10	3	3	0	31	2	9	3	1	0	0
	5	8	0	0	2	1	0	0	26	0	0	0
	6	4	1	2	0	21	5	9	1	0	1	2
	7	3	1	0	1	0	1	0	17	0	0	0

Table F.25: Sparse contingency table showing the interrelation between alignment classes and topic, bin 3 (34-50 posts)

		Topics											
		abortion	*climate change*	*vs capitalism*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*	*legalized prostitution*	*marijuana legalization*	*obamacare*	*women in the military*
Sentiment classes	1	10	0	1	0	2	2	6	5	0	1	1	0
	2	32	1	0	0	26	3	6	31	1	0	0	1
	3	14	1	0	0	28	4	13	8	0	1	2	0
	4	6	0	0	4	0	0	2	25	0	0	0	0
	5	12	0	0	0	1	0	1	37	0	0	0	0
	6	8	1	4	0	45	10	13	1	0	0	0	1

Table F.26: Sparse contingency table showing the interrelation between alignment classes and topic, bin 4 (51-86 posts)

		Topics														
		School Uniforms	*abortion*	*climate change*	*communism vs capitalism*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gays in the military*	*gun control*	*immigration*	*legalized prostitution*	*marijuana legalization*	*obamacare*	*women in the military*
Sentiment class	*1*	0	3	0	0	0	7	8	0	0	0	0	1	0	1	0
	2	0	11	1	0	0	11	1	2	0	5	0	0	1	0	0
	3	0	3	0	0	7	0	0	0	0	21	0	0	1	0	0
	4	0	9	0	0	0	1	0	2	0	4	0	0	0	0	0
	5	1	7	0	1	0	17	3	6	0	1	0	0	0	1	0
	6	0	9	0	0	0	0	0	2	0	10	0	0	0	0	0
	7	0	8	1	0	0	15	1	8	0	3	0	0	0	0	1
	8	0	10	1	0	0	2	0	1	1	5	1	0	0	0	0

Table F.27: Sparse contingency table showing the interrelation between alignment classes and topic, bin 5 (87-136 posts)

		Topics							
		School Uniforms	*abortion*	*climate change*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*
Sentiment classes	*1*	0	1	0	0	0	0	1	6
	2	1	1	1	0	8	1	1	0
	3	0	7	0	0	0	0	1	5
	4	0	2	1	0	6	1	4	0
	5	0	8	0	0	2	1	2	2
	6	0	1	0	1	0	0	0	4

Table F.28: Sparse contingency table showing the interrelation between alignment classes and topic, bin 6 (137-186 posts)

		Topics						
		abortion	*evolution*	*existence of God*	*gay marriage*	*gun control*	*marijuana legalization*	*vegetarianism*
Sentiment classes	*1*	0	0	0	2	0	0	0
	2	3	0	0	2	2	1	1
	3	0	1	0	0	0	0	0
	4	1	2	1	1	1	0	0
	5	1	1	0	0	4	0	0
	6	3	7	1	3	0	0	0
	7	2	2	0	2	0	0	0

Table F.29: Sparse contingency table showing the interrelation between alignment classes and topic, bin 7 (187-236 posts)

		Topics						
		abortion	*climate change*	*death penalty*	*evolution*	*existence of God*	*gay marriage*	*gun control*
Sentiment classes	*1*	2	0	0	5	0	0	0
	2	0	0	0	1	2	0	0
	3	3	1	0	2	0	0	0
	4	3	0	0	0	0	0	0
	5	0	0	1	0	0	0	3
	6	0	0	0	0	0	1	2

Table F.30: Sparse contingency table showing the interrelation between alignment classes and topic, bin 8 (237-286 posts)

Appendix G

Illustrating examples

G.1 Sentiment classes fitted by “gun control”

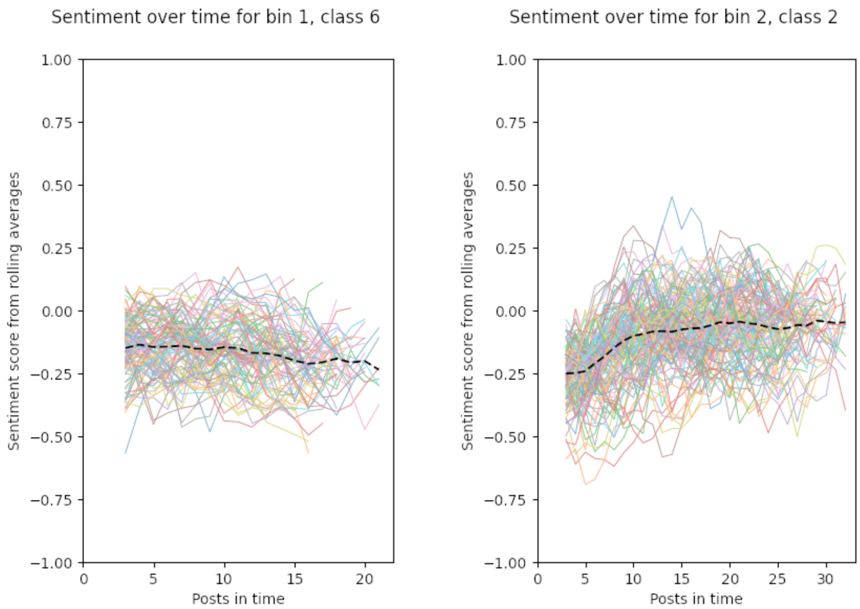


Figure G.1: Sentiment classes which mostly consist of discussions with the topic “gun control”, part 1

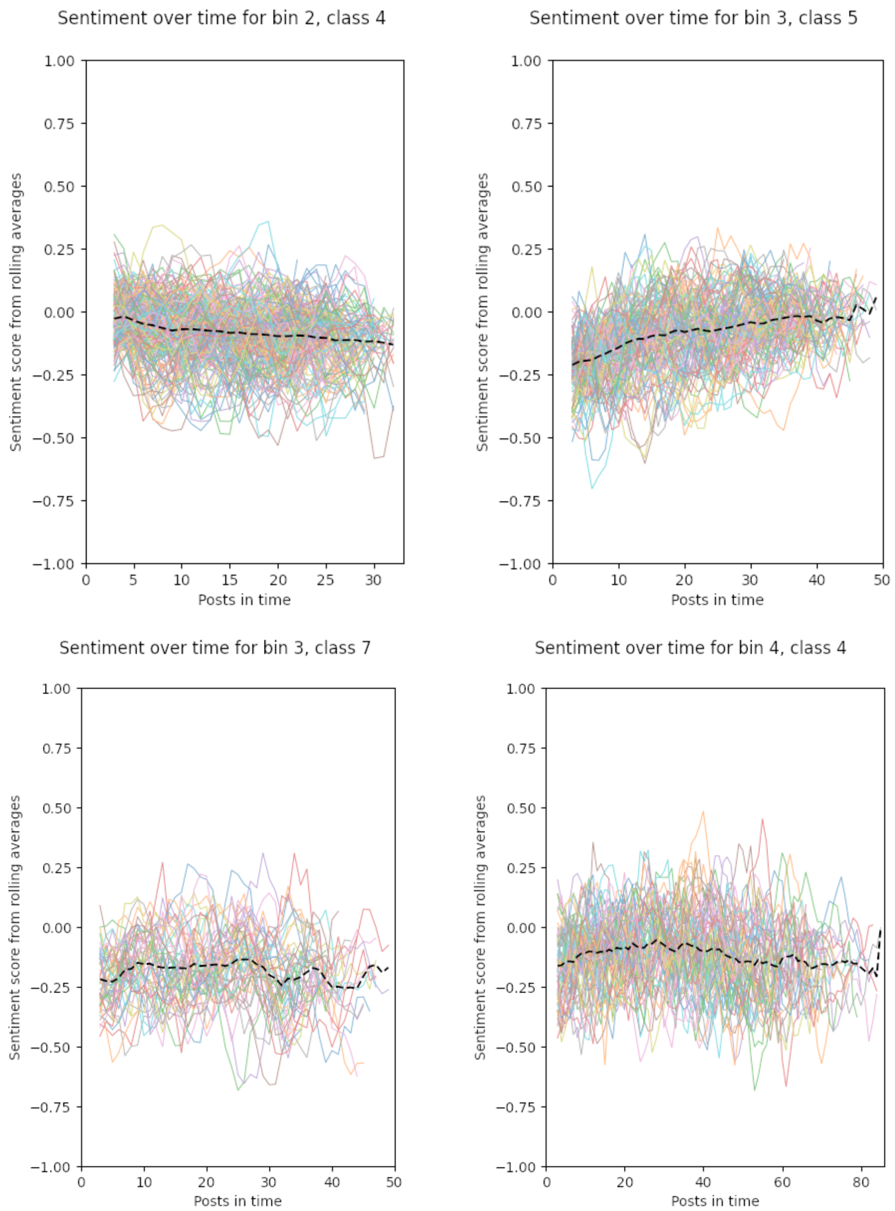


Figure G.2: Sentiment classes which mostly consist of discussions with the topic "gun control", part 2

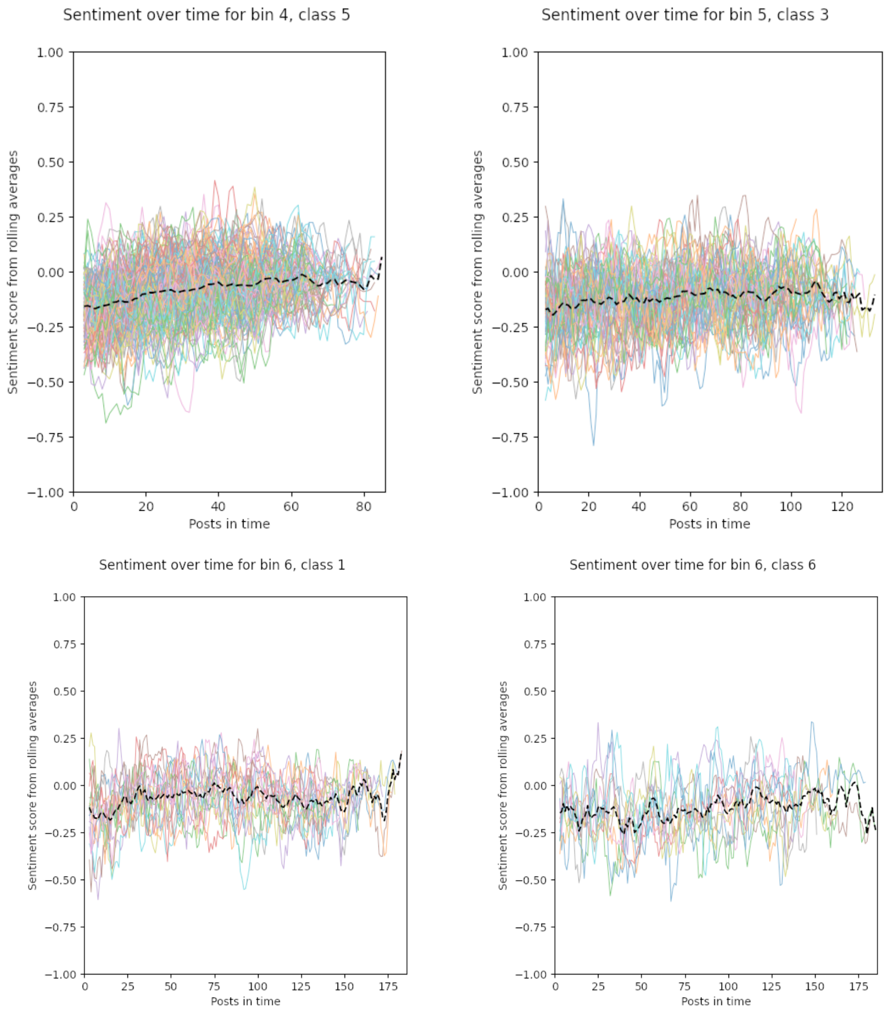


Figure G.3: Sentiment classes which mostly consist of discussions with the topic "gun control", part 3