Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Marius Valen

Context versus Content

A context analysis of AiBA chat data using USE and SBERT

Master's thesis in Information Security
Supervisor: Patrick Adrianus Bours
Co-supervisor: Sushma Venkatesh
December 2023

NTNU
Norwegian University of
Science and Technology

Marius Valen

Context versus Content

A context analysis of AiBA chat data using USE and SBERT

Master's thesis in Information Security
Supervisor: Patrick Adrianus Bours
Co-supervisor: Sushma Venkatesh
December 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Context versus Content: A context analysis of AiBA chat data using USE and SBERT

Marius Valen

December 2023

Abstract

In investigating criminal cases that deal with sexual abuse of children in the form of text, images and videos, it is always challenging to sort this data out of a larger amount of data. With the technological progress in society, the number of devices and the amount of data seized in criminal cases is increasing rapidly. Data storage and the amounts of data has increased exponentially in recent years and it does not seem to be stopping anytime soon. It is becoming more and more demanding to go through these amounts of data and to be able to effectively identify the data that illuminates the criminal relationship. This thesis will explore the sentimentation of messages as an aid to reveal the meaning of the content. We will further see how this scores in different sentence-models. This can lead us to the development of new lexicons for the sentiment of words that can be identified as sexual grooming and new methods to identifying sexual grooming faster and more reliable. This thesis will explore this by using a data set of message data from AiBA and analyze this through the use of Universal Sentence Encoder (USE) and Sentence-Bidirectional Encoder Representations from Transformers (SBERT) and then compare the results from the two models. We found that the sentences score quite differently even though they are contextually identical. This implies that further research to train the language models is needed.

Sammendrag

I etterforskning av saker som omhandler seksuelle overgrep mot barn i form av tekst, bilder og videoer er det alltid utfordrende å sortere denne informasjonen ut fra en større mengde med data. Med den teknologiske fremgangen i samfunnet øker det med både antall enheter og hvor mye data som blir beslaglagt i straffesaker. Datalagring og mengden data har økt eksponensielt de siste årene og det ser ikke ut til å stoppe med det første. Det blir stadig mer krevende å gå gjennom disse mengdene med data og på en effektiv måte kunne identfisere de dataene som belyser det straffbare forholdet. Ved å kunne dele opp chat-meldinger i 'sentiment' kan man sannsynliggjøre hva meningen i innholdet er. Deretter vil vi sammenlikne setningene og se hvordan de scorer i forskjellige setnings-analyser. Dette kan føre oss til utvikling av nye leksikon for å kategorisere ord som kan identifiseres som seksuell grooming og nye metoder for å identifisere seksuell grooming raskere og mer pålitelig. Denne oppgaven vil utforske dette ved å bruke et datasett med meldings-data fra AiBA og analysere denne gjennom bruk av Universal Sentence Encoder (USE) og Sentence-Bidirectional Encoder Representations from Transformers (SBERT) og for så å sammenlikne resultatene fra disse to modellene. Vi fant ut at setningene scorer ganske forskjellig selv om de er kontekstuelt identiske. Dette innebærer at det trengs ytterligere forskning for å trene språkmodellene.

Contents

Figures

Tables

Acronyms

BERT Bidirectional Encoder Representations from Transformers. 5

CSAM Child Sexual Abuse Materials. 1

EDA Exploratory Data Analysis. vii, 2, 3, 8, 17

NLP Natural Language Processing. 5

SBERT Sentence-Bidirectional Encoder Representations from Transformers. iii, v, vii, viii, xi, 2, 5, 6, 10, 11, 15, 19–21

USE Universal Sentence Encoder. iii, v, vii–ix, xi, 2, 5, 11, 14, 19–21

VADER Valence Aware Dictionary and sEntiment Reasoner. 19

Chapter 1

Introduction

In the online space there is a lot of communication between people using chat. Some times these chats are between an adult and a child. Sexual grooming in chat is a serious problem that is important to identify. Studies show that almost 1 in 5 children has been sexually solicited by an adult in the online space[5].

1.1 Motivation

My motivation for this thesis comes from experience with reading chat and messages in real criminal cases. My work as a police superintendent on computer forensics has given me insight and experience in working on cases containing sexual grooming and Child Sexual Abuse Materials (CSAM). I have several times experienced difficulty identifying if the chat is sexual grooming or not. Even though the chat isn't explicitly sexual it is difficult to determine if the chat is grooming or not. And to be able to differentiate and early identify sexual grooming is very important from both a legal certainty and information security point of view. My hope in writing my master thesis is to contribute to earlier detection and hopefully prevention of sexual grooming in the online space.

Analysis of chat data can in many cases be important to understand the underlying meaning of the communication between predator and a child. And exploration and a deeper understanding of the different ways chat data is related can be very important to my line of work within law enforcement.

1.2 Information Security

In our thesis we will explore a specific dataset that contains chat data from several games within the online space. Chat data often contains sensitive and private information and must be handled with the appropriate care and considerations regarding this. The dataset that I am exploring is contained within a Microsoft Azure cloud environment and experiments conducted is done within this platform. By storing the data within this environment will ensure that no data can be copied

elsewhere or lost. This also help me by having the data readily available. Access to this data is only given to people working with the data and uses both login credentials as well as multi-factor authorization. This ensures that no third party can access it.

1.3 Research questions

To first understand the dataset that we will be working on we first have to look at the data itself. By Conducting an Exploratory Data Analysis (EDA) is one way to look at the dataset and to identify the relationships within the dataset. [6]

RQ1 - Can we identify relationships between users, dates, times and sentimentation of the chat messages in the dataset?

The goal here is to find methods to identify relationships within the chat messages in the dataset. This can prove useful for further analysis of the content and context of the chat messages in the dataset.

RQ2 - Will chat messages with the same context, but different content score differenlty or the same when compared?

After conducting an EDA of the dataset we will explore the relationships between the different data. We will then use these relationships to conduct a Universal Sentence Encoder (USE) analysis and compare these results with a Sentence-Bidirectional Encoder Representations from Transformers (SBERT) analysis. This is to potentially identify different scoring when it comes to the same context with different content.

1.4 Thesis structure

This thesis will consist of four parts. The first part will be a literature study of what sexual grooming is and what context and content analysis is. Second part will consist of the analysis of the data. We will there explore models used on the AiBA chat dataset. The last parts will contain a discussion of and conclusion to the findings.

Chapter 2

Literature study

2.0.1 Exploratory Data Analysis

EDA is the method to investigate data sets and summarize their main characteristics. The data is often employed with data virtualization methods. [6]. EDA can be used to look at the data before making any assumptions. It helps the research by identifying obvious errors and can find relations among the variables that can be interesting. [6]

To make it easier to conduct an EDA is to divide the process into six steps. These steps are[12]:

- Observe your dataset
- Find any missing values
- Categorize your values
- Find the shape of your dataset
- Identify relationships in your dataset
- Locate any outliers in your dataset

2.0.2 Sexual Grooming

To first understand what sexual grooming is we have to have a definition of what sexual grooming is. According to Crowell there has been a lack of definition which leads to "miscommunication accross individuals, organizations, and fields in general" (p. 38). [2] As a result they have made a definition that should cover all fields and problems:

> "Sexual grooming is the deceptive process used by sexual abusers to facilitate sexual contact with a minor while simultaneously avoiding detection. Prior to the commission of the sexual abuse, the would-be sexual abuser may select a victim; gain access to and isolate the minor; develop trust with the minor and often their guardians, community, and youth-serving institutions; and desensitize the minor to sexual content and physical contact. Post-abuse, the offender may use

> maintenance strategies on the victim to facilitate future sexual abuse and/or to prevent disclosure." (p. 47) [2]

This definition will be the basis for our further discussion about sexual grooming.

By looking into existing work on online sexual grooming we found that several used data from Perverted Justice [2]. We found this data to be good, but it is a concern that the data is based on chat between predators and members of law enforcement. Law enforcement can have a goal with the chat they are conducting and not to be presentable as "real" chat with minors. We will be basing the data for this thesis on data from the AiBA database.

2.0.3 Context based classification and language models

Context based classification has been done a lot on X/Twitter as they have looked into sentiment classification which extracts both the context and content. [11] This way of looking at tweets use both conversation-based context, author-based context and topic-based context. The challenge with using this model for our thesis is that it uses X/Twitter's sentiment classifications to determine the context. Their result however was that a context-based neural network model had improved performance compared to other models[11].

Other papers have also explored if context really matters at all when it comes to toxicity detection. [9] The interesting part with this analysis is if the previous context (message) matters for how toxic the next message is. One of the research questions was that

> does context improve the performance of toxicity classifiers, when they are made context-aware? [9]

Their conclusion was that context has a statistically significance when looking at toxic messages. This could be valuable data to look at when looking at contextual classification of sexual grooming[9].

2.0.4 Content

My experience with working with chat data and reading chat data in criminal cases is that I have had several experiences where context is difficult to determine. For instance, In a case I worked on where a minor had been selling naked pictures to adults, I read a whole chat between an adult and a child where the adult sent her several small amount of money and very little context. With the amount transfered there was attached heart and smiley emojis. But when I looked into who this adult was it turned out to be her father. The whole context flipped over because of this. It is normal for a father to send money to their child and also send smiley and heart emojis. This underlines how important context is when analysing chat data.

2.1 Different sentence comparison models

2.1.1 Universal Sentence Encoder (USE)

Universal Sentence Encoder is a model that encodes sentences into high-dimensional vectors. [8] The vectors can then be used as text classification, semantic similarity, clustering and more in various Natural Language Processing (NLP) tasks. This type of model is also optimized and trained for sentences, short paragraphs or phrases. The model makes it easier to do sentence embeddings. This makes it a suitable model for use in our analysis. This model is available for free as a part of the Tensorflow hub [13]. Similarity using USE is calculated from a score between 0.0 to 1.0 where as 1.0 is 100% similarity and 0.0 is 0% similarity.

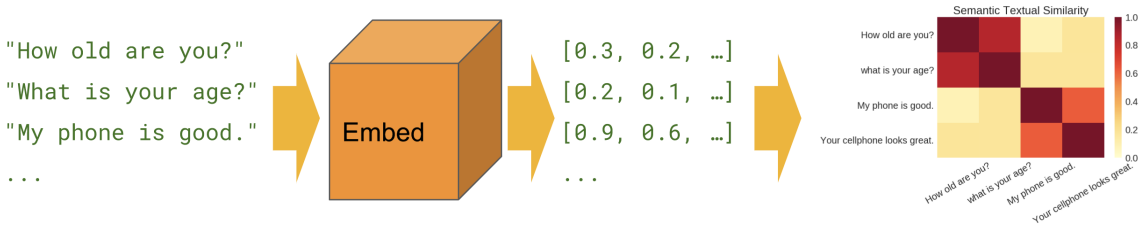


Figure 2.1: Example of USE in use with sentence-comparison[8]

As shown in figure 2.1[8] we can see a how USE is embedding the sentences and scoring them from 0.0-1.0 based on semantic textual similarity[8].

2.1.2 Sentence-Bidirectional Encoder Representations from Transformers (SBERT)

Sentence-BERT or SBERT is a modified pre-trained BERT network that is used to derive semantically meaningful sentence embeddings[10]. When compared to Bidirectional Encoder Representations from Transformers (BERT) it drastically reduces the effort and time for most similar pairs from 65 hours with BERT to about 5 seconds using SBERT[10]. The similarity using SBERT is calculated with a score between -1 to 1 whereas -1 is least similarity and 1 is 100% similarity.[10]

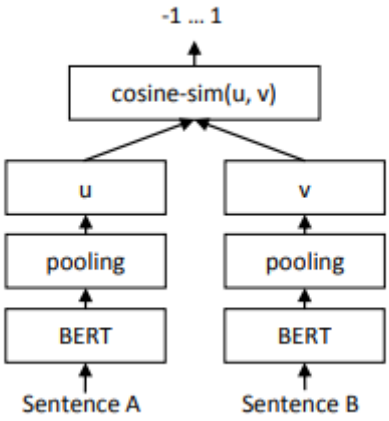


Figure 2.2: SBERT architecture to compute similarity scores[10]

The figure from N. Reimers and I. Gurevych shows how the SBERT architecture is when it comes to similarity comparison between two sentences[10]. This is the same model we will be using to compare one sentence to another sentence from the AiBA dataset.

Chapter 3

Analysis of the dataset

3.1 Dataset

The dataset we are using in this thesis is provided by AiBA and consists of chat data from several online games that involves communications between users. This dataset is localized in a closed Azure environment and will not be provided as an appendix.

Before running the first data we have to see how the chat data is structured. The chat data consists of over 28 million chat messages and were divided into several Apache Parquet files. These files is an open source, column-oriented data file format that is designed for efficient data storage and retrieval.

The chat data was structured into the following columns:

Parquet dataset							
dateUtc	messageID	context	gameId	initiator	receiver	content	studentid

Table 3.1: Structure of the dataset from AiBA

- 'dateUtc' consists of the date as well as timestamp for the message.
- 'messageId' consists of the unique identifier for that message
- 'context' consists of the unique identifier for that chat conversation
- 'gameId' consists of the unique identifier for which game the chat was from
- 'initiator' consists of the unique identifier of the sender of the message
- 'receiver' consists of the unique identifier of the recipient of the message
- 'studentid' consists of the unique number for which student who has access to the data

The dataset contains 28.428.425 lines of chat data. This data consists of chat data ranging from the time 00:00:00 on the 1st of August 2022 to the time 23:59:59 on the 31st of October 2022.

By understanding how the dataset is structured we can continue exploring the content of the dataset.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Relationship to dates

When looking into the data the first date in UTC present in the dataset is 2022-08-01 and the last date 2022-10-31.

When analysing which dates has the most frequent chat it is the following dates (the five most frequent dates):

Date	Frequency
2022-08-01	452897
2022-08-09	449813
2022-08-03	446225
2022-08-02	445580
2022-08-04	443435

Table 3.2: Dates with highest frequency of chat messages

3.2.2 Relationship to weekdays

When sorting the data and looking into the relationship between frequency of messages and certain days of the week the data has the following relationship sorted by highest frequency to lowest:

Weekday	Frequency
Sunday	4587048
Monday	4438976
Saturday	4144138
Tuesday	3988970
Wednesday	3815218
Friday	3730971
Thursday	3723104

Table 3.3: Weekdays with highest frequency of messages

When put into a histogram the data shows clearly that there are the most activity on Sundays:

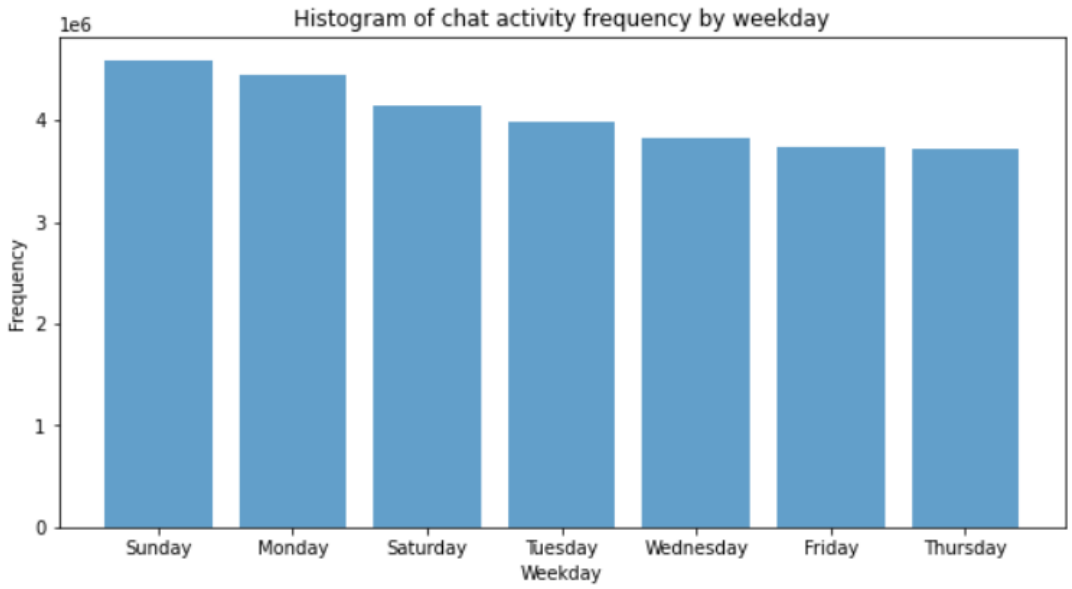


Figure 3.1: Frequency of messages sorted on days of week

This is somewhat understandable as most people have the day off on Sundays. This information could prove to be useful for the moderators of the different chat spaces. When should they have most moderators on duty?

3.2.3 Which initiators has the highest frequency?

In the data the users have their own unique hexadecimal value. We can use this value to look deeper into each user's data.

When looking deeper into the chat data we wanted to see who is sending the most chats? By using EDA we could find that the top five senders are the following:

User ID	Chat messages
3860DB4F2C5A4E17484EBED553AFE685	56415
6CB2B06ADE5291093215E88C5BF6F027	54264
67E094B483937DF7F0C9D733DF6EBAB3	50138
CC92570924127FB82FE8257115FCADEB	39918
4F68AEF24938E0BE4B4C8BED89C4FB3D	37339

Table 3.4: Initiators with highest fequency of messages

When put into a histogram it gives us the following and clearly shows that the user with the ID '3860DB4F2C5A4E17484EBED553AFE685' is the most active sender of chat messages with a total of 56415 chat messages:

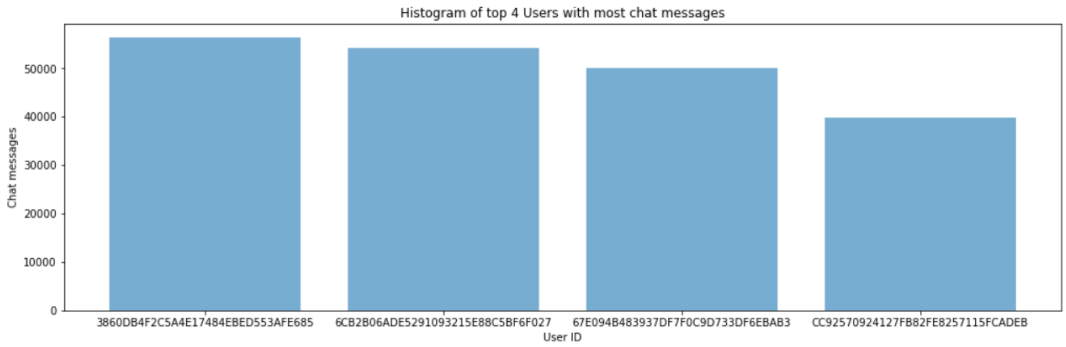


Figure 3.2: Top four users initiating chats

By identifying the users who have the highest frequency of initiating chats we can more accurately know who to moderate and look into what these users are sending. Is there a certain reason for why they are initiating this many chats? Are they spamming the game they are connected to? These are interesting findings that can be useful for moderation of the chats.

Who receives the most chat messages?

By using the unique user ID we can also identify which of the users receive the most chat messages. By using EDA and listing the five most frequent receivers we can find the following data:

User ID	Chat messages received
6CB2B06ADE5291093215E88C5BF6F027	56825
3860DB4F2C5A4E17484EBED553AFE685	54140
67E094B483937DF7F0C9D733DF6EBAB3	50832
0106008BDE562EFF03BE04E139246054	42190
CC92570924127FB82FE8257115FCADEB	40649

Table 3.5: Users with most messages received

This analysis clearly shows that the user with the ID '6CB2B06ADE5291093215E88C5BF6F027' is the user in this dataset that receives the most chat messages with a total of 56825 messages.

3.2.4 Sentiment analysis of the dataset

Before we can use the dataset in a SBERT analysis we first have to apply sentiments to each row of chat message. This can be done by using different lexicons to analyse the data and add classifications. When researching this topic I found that there are no "best" lexicon as there are different lexicons serving different purposes. It all depends on the data that is getting a sentiment-analysis. There are several open-source lexicons as well as commercially licensed that cost money

to use. For this thesis I will rely on the "Valence Aware Dictionary and sEntiment Reasoner" or VADER lexicon.

The reason for using this lexicon instead of others is that it is specifically designed for chat messages from social media and text messages. It takes into account the context and syntaxes used in chat messages from social media and text messages. [1] Which both are relevant to the dataset used in this thesis.

By analyzing which chat messages that are positive, negative and neutral we can use the data more specific in a BERT model.

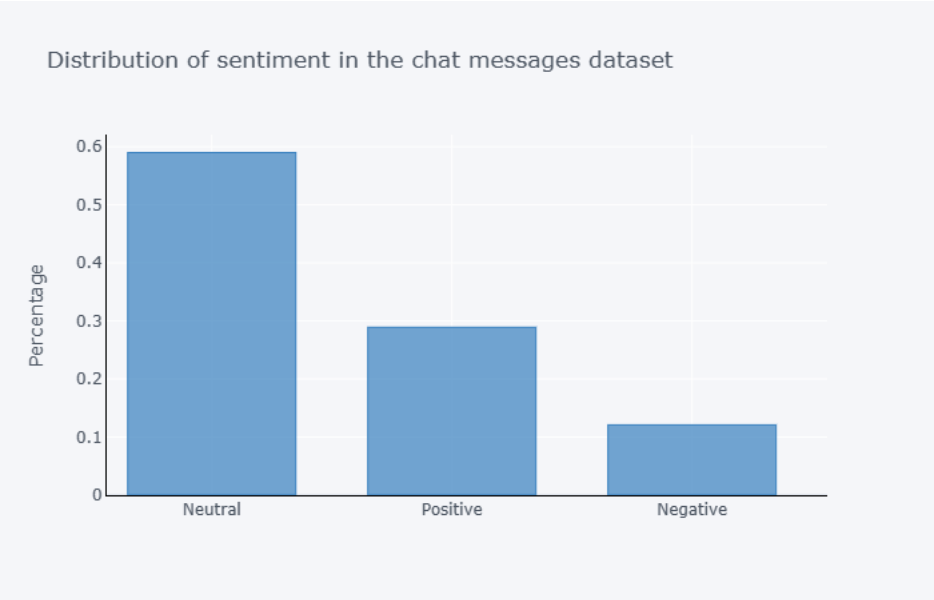


Figure 3.3: Distribution of sentiments

Here we can see that the dataset consists of 59% neutral, 29% positive and 12% negative chat messages. When doing my further analysis of the dataset we are going to focus on the messages that are sentimented as positive. We want to look further to see if there are any difference in how similar chat messages are categorized.

3.3 Digging deeper into the dataset using USE and SBERT

3.3.1 N-grams

By continuing using EDA we can look deeper into how the different messages are sentimented. To do so we have to look into what N-grams are [7]. N-grams are separating how many words in which instance that is constructing the sentence. The larger the value of N the more context the sentence will have. But the larger value of N, then less N-grams will be found as the frequency of a specific N-gram will decrease. [4]

By looking at an example sentence: "This is a sentence", and separating the different words that builds the sentence. It can be separated into unigrams (one word at a time), bigrams (two words at a time) or trigrams (three words at a time). The different N-grams can be formulated like this[7]:

$\mathbf{P}_{\texttt{unigram}}(w_1,\ldots,w_4) = \texttt{P}(\text{"}this\text{"}) \cdot \texttt{P}(\text{"}is\text{"}) \cdot \texttt{P}(\text{"}a\text{"}) \cdot \texttt{P}(\text{"}sentence\text{"})$
$\mathbf{P}_{\texttt{bigram}}(w_1,\ldots,w_4) = \texttt{P}(\text{"}this\text{"}|<s>)\cdot\texttt{P}(\text{"}is\text{"}|\text{"}this\text{"})\cdot\texttt{P}(\text{"}a\text{"}|\text{"}is\text{"})\cdot\texttt{P}(\text{"}sentence\text{"}|\text{"}is\text{"})$
$\mathbf{P}_{\texttt{trigram}}(w_1,\ldots,w_4) = \texttt{P}(\text{"}this\text{"}|<s><s>)\cdot\texttt{P}(\text{"}is\text{"}|<s>,\text{"}this\text{"})\cdot\texttt{P}(\text{"}a\text{"}|\text{"}this\text{"},\text{"}is\text{"})\cdot$
$\texttt{P}(\text{"}sentence\text{"}|\text{"}is\text{"},\text{"}a\text{"})$

This can more easily be shown in a more structured way [7]:

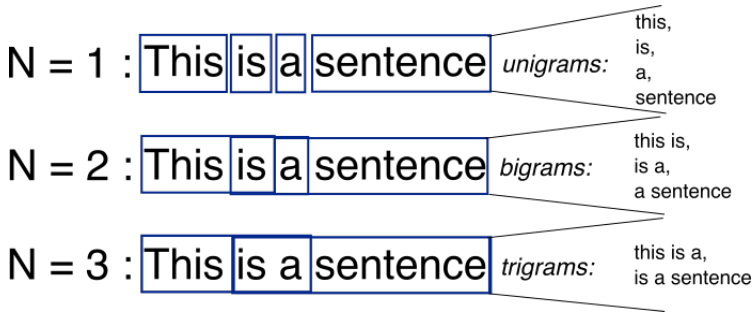


Figure 3.4: N-gram structure explained [7]

We can use the different Unigrams, Bigrams and Trigrams to look into which words/sentences are most used in our dataset of chat data. We will use the results of the positive results for the scope of this study:

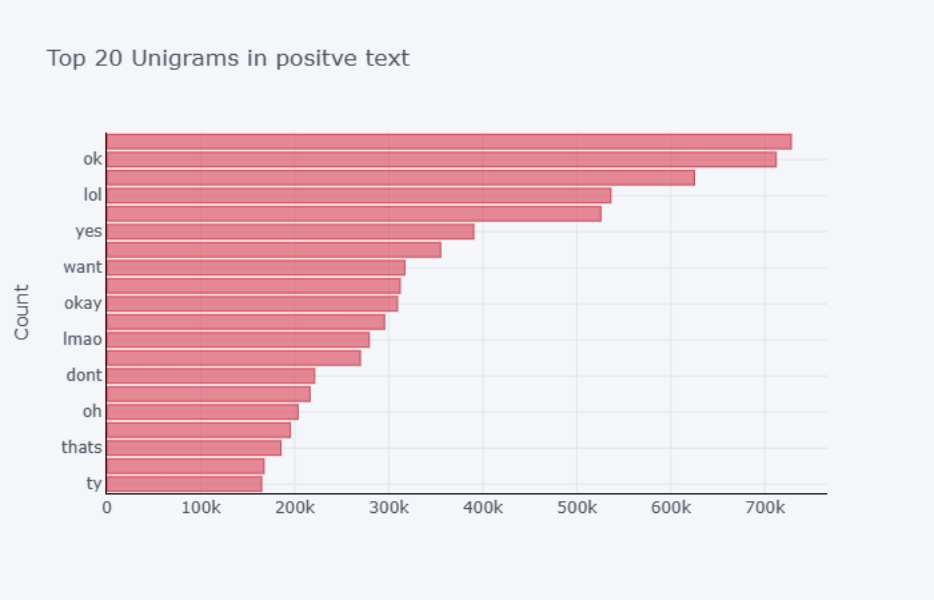


Figure 3.5: Most used positive unigrams

Here we can see the most used positive words. As our N-gram was set to One (1) then it will only identify the single positive word.

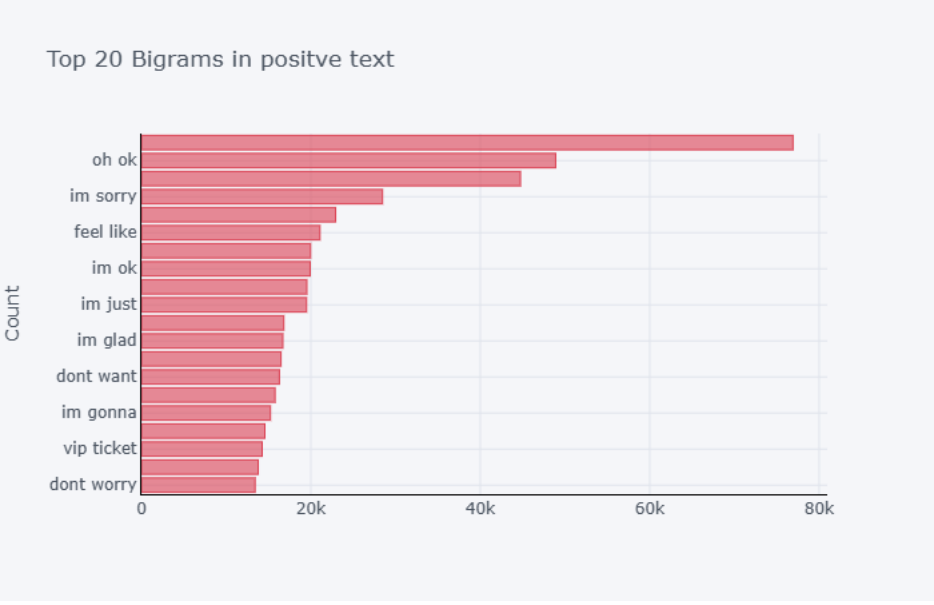


Figure 3.6: Most used positive bigrams

By increasing our N-gram to two (2) we start to get shorter messages consisting of more content. For example: "oh ok", and "im glad".

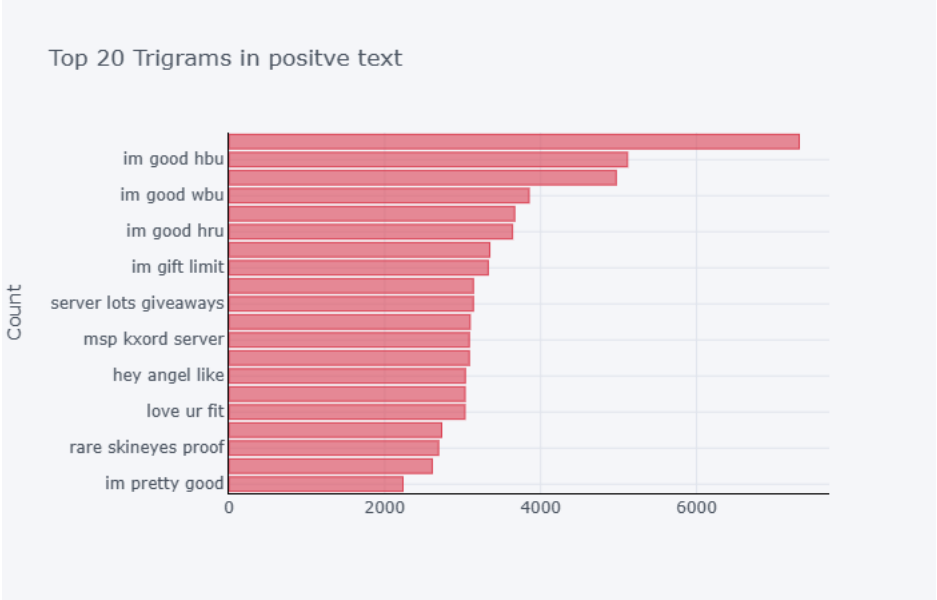


Figure 3.7: Most used positive trigrams

And finally by increasing N-gram to three (3) we get complete, shorter sentences that we will use to further investigate if the same context, but different content will score similar in further analysis. When looking at the trigrams we can see that there are several sentences that we can try to compare against each other to see how they score in different models. For instance the sentence: "im good hbu" (I'm good how about you?), "im good hru" (I'm good how are you?) and "im good wbu" (I'm good what about you?) all have different content, but the context is the same. For the scope of this analysis we will rely on these sentences as examples for further studies.

3.3.2 Universal Sentence Encoder (USE)

We want to try these different sentences with the same context into different methods of comparing them against each other. One method of comparing sentences is the Universal Sentence Encoder (USE) [3]. This model is a way to compare the similarity of sentences and scoring them in a score table from 0.0 to 1.0.[3]

When comparing the different sentences in a USE model they do score a bit different. On a scale where 0.0 is completely different and 1.0 is exactly the same[3] the different sentences score were this:

Sentence 1	Sentence 2	USE Score
"im good hbu"	"im good hru"	0.895
"im good hbu"	"im good wbu"	0.92
"im good hru"	"im good wbu"	0.875

Table 3.6: USE score of the different positive trigrams

3.3.3 Sentence-Bidirectional Encoder Representations from Transformers (SBERT)

When running the same sentences as well as the sentences without the abbreviations in a SBERT model it gave the some interesting scoring. These scores are from a range from -1 to 1 [10]:

Sentence 1	Sentence 2	SBERT Score
"im good hbu"	"im good hru"	0.6139
"im good how about you"	"im good how are you"	0.7908
"im good hbu"	"im good wbu"	0.6630
"im good how about you"	"im good what about you"	0.9064
"im good hru"	"im good wbu"	0.6872
"im good how are you"	"im good what about you"	0.7369

Table 3.7: SBERT score of the different positive trigrams

As seen in table 3.7, we can see that the scoring when using abbreviations are quite different from the sentence meaning. The sentences in general scores higher on a similarity-scale when abbreviations are taken out of the equation.

Chapter 4

Discussion

4.1 EDA Analysis

When analysing the dataset using EDA we could find several interesting key findings. The dataset is a very limited set of data when it comes to how few dates it contains, given that there are over 28.4 million messages in a time-span from the 1st of August 2022 to 31st of October 2022. It is only a small portion of the total chat data that the different games/applications contains. So the same analysis won't necessary look the same when conducting it on other parts of the complete chat dataset.

4.1.1 Relationship to time and dates

We analyzed the dataset based on time and date. What we found was that the day with highest frequency of chat messages was Sunday. While Monday and Saturday was runner up. This was very interesting to be able to pinpoint the day of the week that has the most user interactions. By knowing this we can identify which day where most monitoring and moderation is needed.

4.1.2 Sentiment analysis

After analysing the data and placing every chat message in to a sentiment we found that the majority of the chat messages was of a neutral nature, a whole 59% of the chat messages. The distribution for positive was 29% while the negative chat messages was 12%.

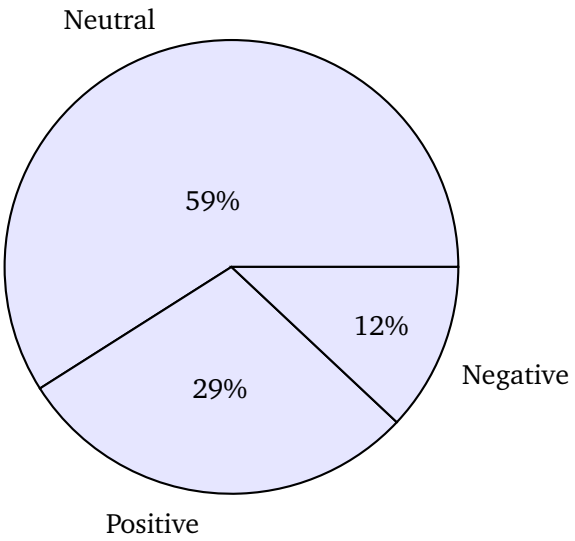


Figure 4.1: Distribution of sentiment in the dataset

By separating the different sentences into sentiments we can further split the data into positive, negative and neutral meaning. This makes the amount of data less stressful on the different language models. In our testing we used the VADER lexicon[1]. But if a lexicon is made for words that can identify sexual grooming, then sentimentation like this can be very useful to identify sexual grooming from a larger dataset.

4.2 N-grams analysis

When looking at the results of the N-gram analysis of the chat data we could identify sentences that are often used that has the same context, but different content. This will be a challenge when working with chat data that contains many variables with both abbreviations as well as emojis. But looking at the most common positive unigram-words used was:

- ok
- lol
- yes
- want
- okay
- lmao
- dont
- oh
- thats
- ty

Table 4.1: Most used positive unigrams

In the table 4.1 we can see that single words that makes a positive statement. These words are present in the VADER lexicon and directly found in the chat data. Several of these also overlap, like 'ok' and 'okay' as well as 'lol' (laughing out loud) and 'lmao' (laughing my ass off). Where as these findings are interesting they do not make up a whole sentence as they are most likely single words taken out of a sentence. That is why we increased N-gram number to 2 and 3 to find the positive statements in bigrams and trigrams.

While looking into the trigrams we could identify several sentences that has the same context, but different content. These sentences were:

- im good hbu
- im good wbu
- im good wbu
- im good hru

Table 4.2: Most used positive trigrams

The sentences from table 4.2 was very useful in our further analysis. We limited the analysis to these sentences for the scope of this study and could further use these in our USE and SBERT analysis. It is quite interesting to see such similarities in a dataset of over 28 million lines of chat messages. By being able to use the VADER lexicon and find such similarities then this can be used the same if we had a lexicon for sexual grooming. In this way we could localize the occurrence of sexual grooming.

4.3 USE findings

We compared the same contextual messages with both USE and SBERT and it gave us an understanding on how the two models differentiate the same messages. When compared in a USE analysis the sentences scored 0.895, 0.92 and 0.875 on a scale from 0 to 1. We find these sentences to be very close to each other

and they score very similar. But then again they are not identified as 1 to 1 in sentence meaning. Showing that the models do not identify the context as the same. This shows a weakness in the different models in regards to same context of the different sentences. This might imply that more training of the models is needed to make them more accurate. This implies also that models to be trained to identify sexual grooming is needed.

4.4 SBERT findings

By comparing the same sentences from figure 3.7 that were used in the USE analysis the results was quite different. When just looking at the sentences without the abbreviations the results was:

Sentence 1	Sentence 2	SBERT Score
"im good hbu"	"im good hru"	0.6139
"im good hbu"	"im good wbu"	0.6630
"im good hru"	"im good wbu"	0.6872

Table 4.3: SBERT score of the different positive trigrams with abbreviations

As we can see the four different sentences score very similar and are recognized not far from each other. All though they don't score a 1 they are in the upper region in a score range of -1 to 1. But when we remove the abbreviations and insert the meaning of the abbreviations they score much higher:

Sentence 1	Sentence 2	SBERT Score
"im good how about you"	"im good how are you"	0.7908
"im good how about you"	"im good what about you"	0.9064
"im good how are you"	"im good what about you"	0.7369

Table 4.4: SBERT score of the different positive trigrams without abbreviations

As seen in table 4.4 we see that they score significant higher. It is interesting to see that the sentences "im good hru" and "im good wbu" scored higher (0.6872) than "im good hbu" and "im good wbu" (0.6630) with abbreviations. While the former sentence scores 0.7369 and the latter scores a 0.9064. This shows that abbreviations and how they are interpreted is a challenge when it comes to using these models for analyzing sentences and their content and context. In this instance there will be a challenge to identify and interpret every abbreviations, emojis and hidden meanings that lies behind the language of sexual grooming. By developing a lexicon to contain these types of abbreviations can provide a background for training the sentence-models to identify sexual grooming in the chat data.

Chapter 5

Conclusion and future work

5.1 Conclusion

We will now revisit the research questions and look into the research done in previous chapters to answer these:

RQ1 - Can we identify relationships between users, dates, times and sentimentation of the chat data?

When exploring and analyzing the dataset we could see that there is several different relationships between the different columns of data[6]. We could also find several sentences which was suited to be sentimented[6]. By conducting an Exploratory Data Analysis we could find several interesting relationships within the dataset both related to their content, context and also date and time.

RQ2 - Will chat messages with the same context, but different content score differently or the same when compared?

By adding sentiments to the dataset[6] and combined this with N-grams by taking out the positive messages[7] we could find several sentences that contains the same context, but have different content. By comparing these sentences in both a USE and SBERT we could find that these sentences scores differently to each other although they have the same context/meaning. This analysis also showed that abbreviations in chat messages can make an impact on how the sentences scores when compared to each other[3][10].

To summarize; there are several interesting relationships within the data which later can be used for contextual and content analysis. By scoring the different content, but with the same context we can see that they score different, but not far from each other.

5.2 Future work

For further work we would like to see a more specific lexicon made to sentiment chat messages in the context of sexual grooming. During our research we could not find a lexicon related to this kind of chat. Such a lexicon would prove really helpful in uncovering and investigating sexual grooming of children in the online space. Having a lexicon like this could prove useful and work as a tool to identify sexual grooming in a large dataset like the data from AiBA.

A wider analysis of the data could also be of use to further extend the results on the contextual meaning in the sentences of the same context, but with different content. With this analysis in combination of a sexual grooming lexicon would prove a powerful toolset to discover and further investigate sexual grooming of children in the online space.

Bibliography

[1] cjhutto. *VADER - Sentiment Analysis*. URL: https://github.com/cjhutto/vaderSentiment/tree/master (visited on 30/11/2023).

[2] Charles R. et al. Crowell. *Using Luring Communication Theory to Analyze the Behavior of Online Sexual Offenders*. IGI Global, 2020.

[3] Sheng-yi Kong Nan Hua Nicole Limtiaco Rhomni St. John Noah Constant Mario Guajardo-Cespedes Steve Yuan Chris Tar Yun-Hsuan Sung Brian Strope Daniel Cer Yinfei Yang and Ray Kurzweil. *Universal Sentence Encoder*. URL: https://arxiv.org/abs/1803.11175 (visited on 07/12/2023).

[4] Deepai.org. *N-grams*. URL: https://deepai.org/machine-learning-glossary-and-terms/n-gram (visited on 04/12/2023).

[5] Emily A. Greene-Colozzi. 'Experiences and Perceptions of Online Sexual Solicitation and Grooming of Minors: A Retrospective Report'. In: *Journal of Child Sexual Abuse* 29 (2020), pp. 836–854.

[6] IBM. *What is exploratory data analysis?* URL: https://www.ibm.com/topics/exploratory-data-analysis (visited on 17/11/2023).

[7] Antonio Maiolo. *The Speech Recognition Wiki*. URL: https://web.archive.org/web/20180427050745/http://recognize-speech.com/language-model/n-gram-model/comparison (visited on 04/12/2023).

[8] Dimitre Oliveira. *Universal Sentence Encoder*. URL: https://huggingface.co/Dimitre/universal-sentence-encoder (visited on 11/12/2023).

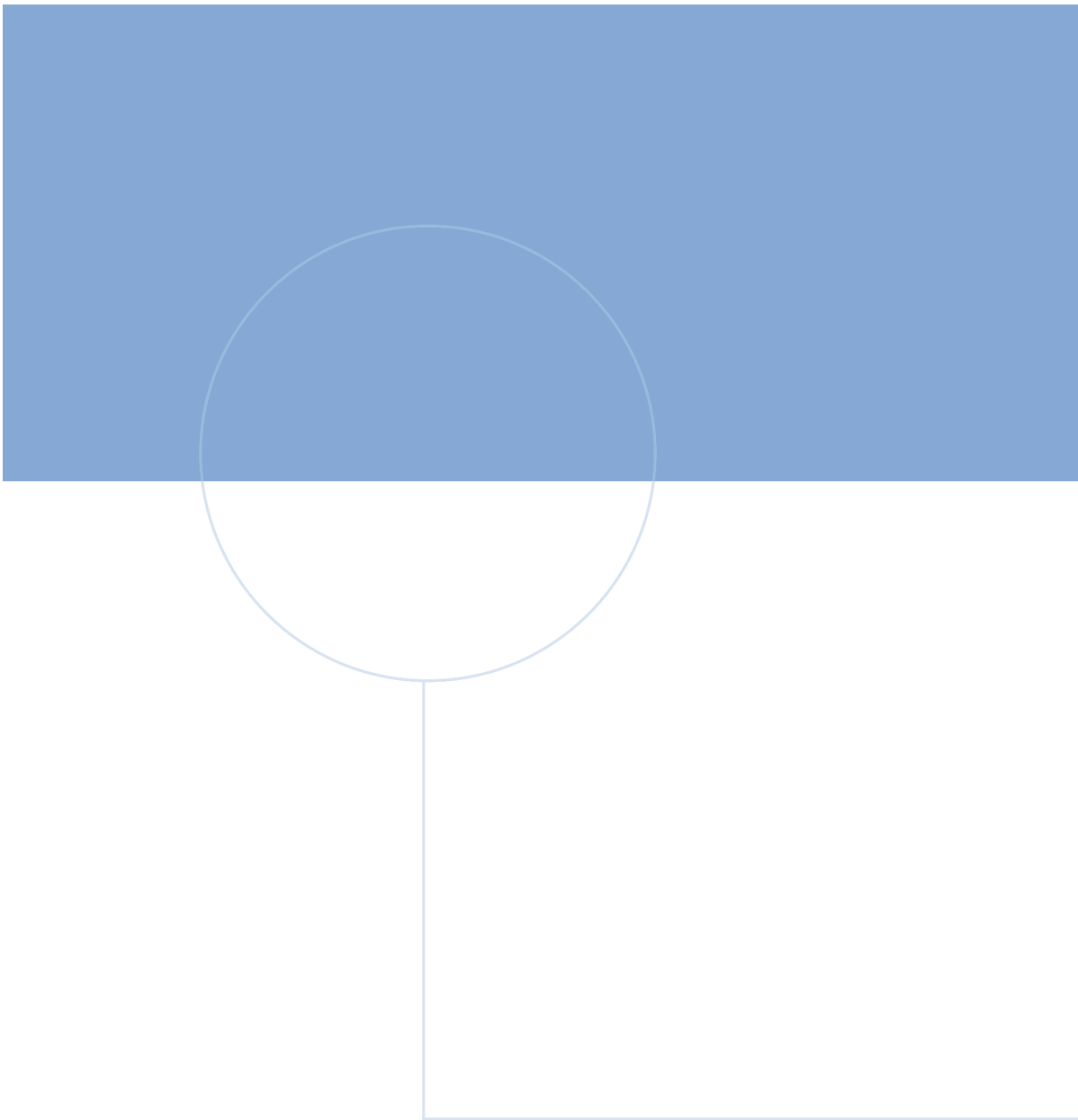
[9] John Pavlopoulos et al. *Toxicity Detection: Does Context Really Matter?* URL: https://arxiv.org/abs/2006.00998 (visited on 21/09/2023).

[10] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. URL: https://arxiv.org/pdf/1908.10084.pdf (visited on 11/12/2023).

[11] Y. Ren et al. 'Context-Sensitive Twitter Sentiment Classification Using Neural Network.' In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (2016).

[12] Indeed Editorial Team. *How To Conduct Exploratory Data Analysis in 6 Steps*. URL: https://www.indeed.com/career-advice/career-development/how-to-conduct-exploratory-data-analysis (visited on 17/11/2023).

[13] TensorFlow. *TensorFlow.org*. URL: https://www.tensorflow.org/ (visited on 11/12/2023).



NTNU

Norwegian University of
Science and Technology