



UNIVERSITÀ
DEGLI STUDI
DI MILANO

LA STATALE

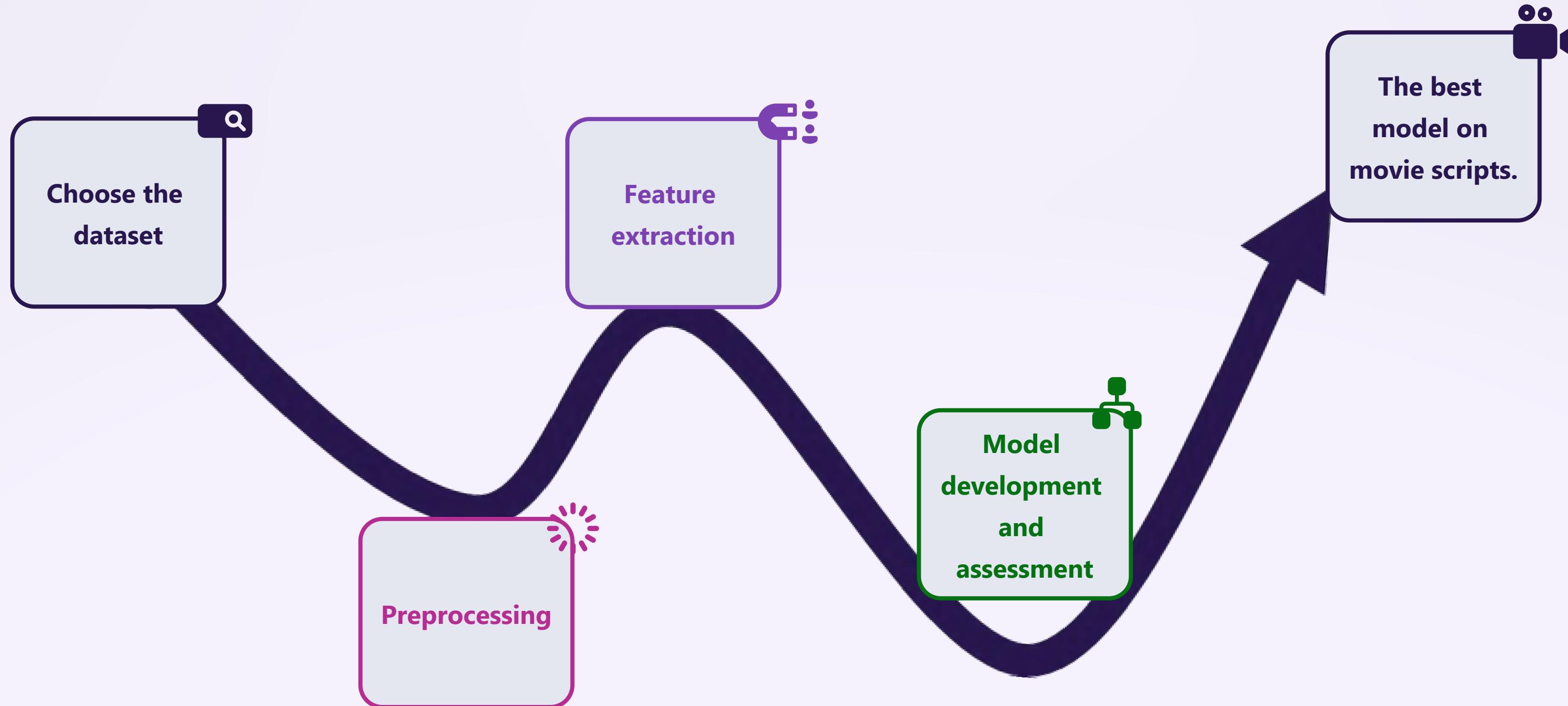


How do you feel, my
dear?



Text Mining and Sentiment Analysis
Vojimir Ranitovic 963780

Map of the project



Dataset

- For this task, the **WASSA-2017** dataset is used.
- It consists of **7102** short texts - tweets.
- Each tweet is labeled by one of **four** labels.

Dataset

- For this task, the **WASSA-2017** dataset is used.
- It consists of **7102** short texts - tweets.
- Each tweet is labeled by one of **four** labels.



Joy



Anger



Sadness



Fear

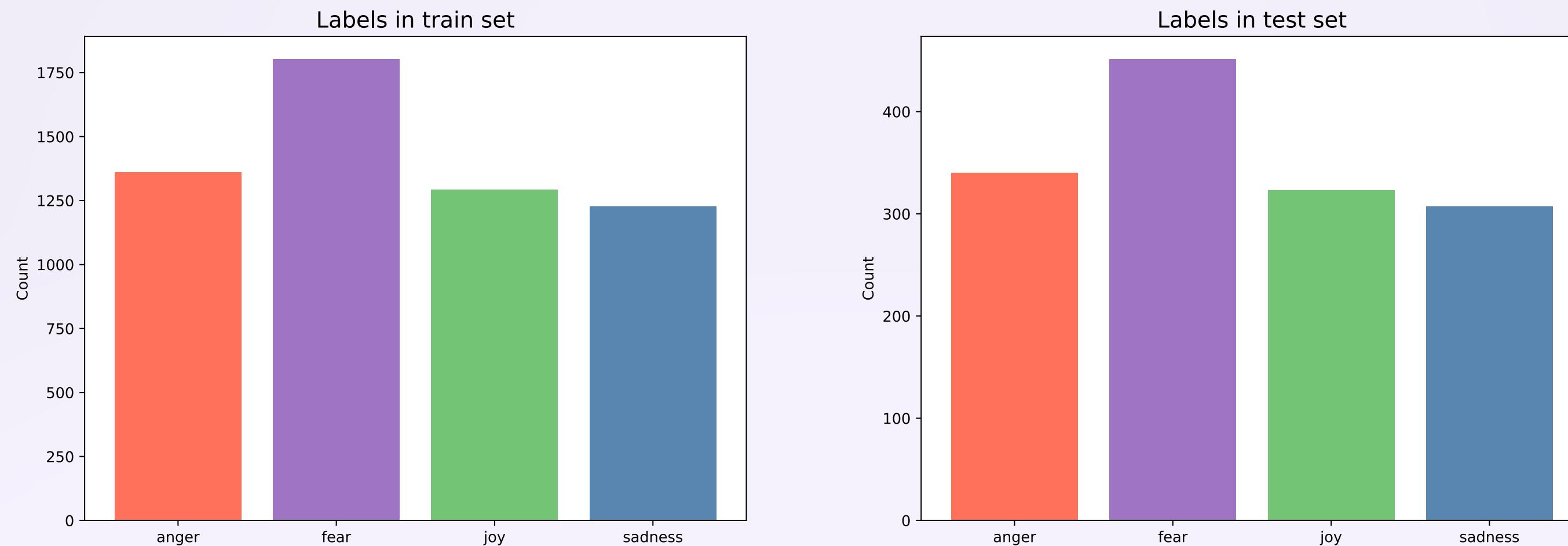
Dataset

Id	Text	Emotion
10000	How the fu*k! Who the heck! moved my fridge!... should I knock the landlord door. #angry #mad ##	anger
10001	So my Indian Uber driver just called someone the N word. If I wasn't in a moving vehicle I'd have jumped out #disgusted	anger
10002	@DPD_UK I asked for my parcel to be delivered to a pick up store not my address #fuming #poorcustomerservice	anger
Id	Text	Emotion
20000	I feel like I am drowning. #depression #anxiety #failure #worthless	fear
20001	I get so nervous even thinking about talking to ***** I wanna die	fear
20002	I lost my blinders #panic	fear
Id	Text	Emotion
30000	Just got back from seeing @GaryDelaney in Burslem. AMAZING!! Face still hurts from laughing so much #hilarious	joy
30001	Oh dear an evening of absolute hilarity I don't think I have laughed so much in a long time! 😂	joy
30002	Been waiting all week for this game ❤️❤️❤️ #cheer #friday ❤️	joy
Id	Text	Emotion
40000	Depression sucks! #depression	sadness
40001	Feeling worthless as always #depression	sadness
40002	Feeling worthless as always	sadness

Data preprocessing

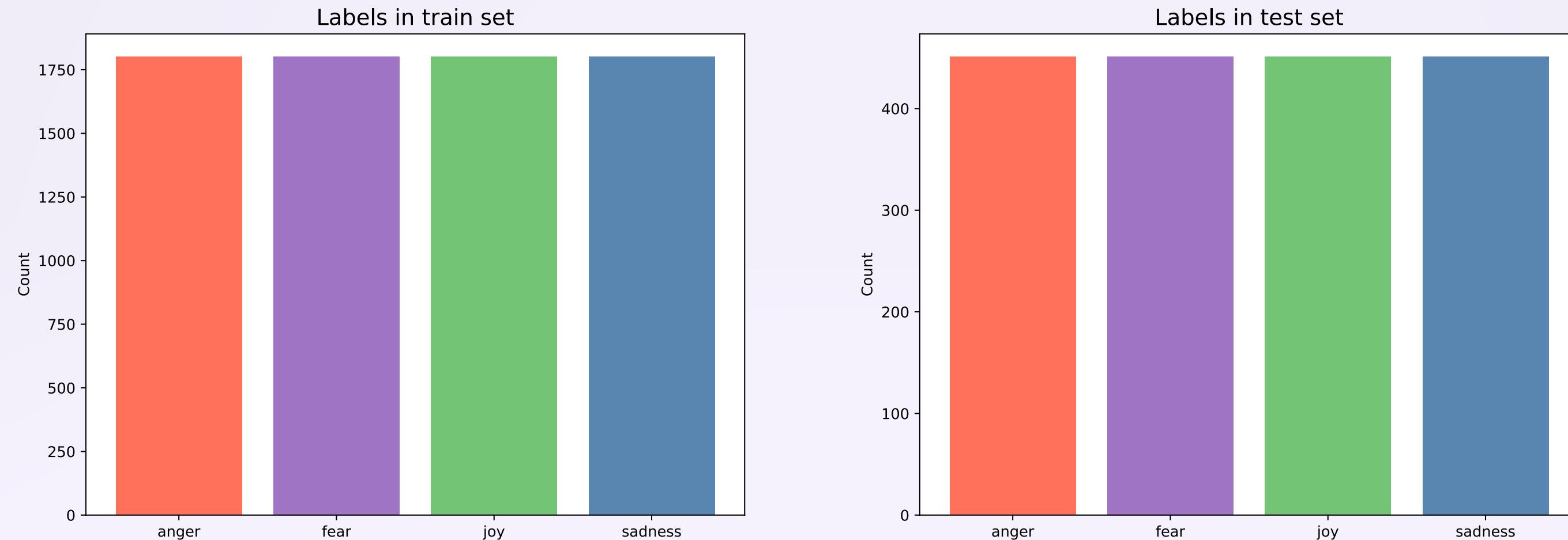


Oversampling



- Stratified splitting into **train** and **test** set.
- Slightly **dominance** by class "fear".
- **Oversampling** other classes until **balance** is reached.

Oversampling*



- Function **RandomOverSampler** just randomly copies minority samples.
- To make newly generated samples more **diverse** → do **synonyms** replacement.

Oversampling

- **Replace_with_synonyms** function **replaces** nouns, adjectives, and adverbs.
- Those **changes** are **only** in the **newly** generated/copied **instances**.
- Original instances from the dataset **remain** the same.

Oversampling

- **Replace_with_synonyms** function **replaces** nouns, adjectives, and adverbs.
- Those **changes** are **only** in the **newly** generated/copied **instances**.
- Original instances from the dataset **remain** the same.

*Goodfellas was authentic and **sort of** provided the anchor for the genre. **Also**, Pesci, DeNiro, and Liotta were perfect **trio**. But, popular **opinion** is that Casino has way **better** ending! @DeNiro that acting was 🔥 🔥 😊 . I want sequel #Casino! 🎉*



*Goodfellas was authentic and **kind of** provided the anchor for the genre. **Besides** , Pesci , DeNiro , and Liotta were perfect **three** . But , popular **impression** is that Casino has way **well** ending ! @DeNiro that acting was 🔥 🔥 😊 . I want sequel #Casino ! 🎉*

Text cleaning

- Removing: punctuation, digits, links, mentions, stopwords, and words with less than three characters.
 - Not removing: words like "no", "nor", "**not**", and "against".
 - Not removing: **hashtags** and **emojis!**
 - Lemmatization



smile face smile eye



loudly cry face

Text cleaning

Goodfellas was authentic and kind of provided the anchor for the genre. Besides , Pesci , DeNiro , and Liotta were perfect three . But , popular impression is that Casino has way well ending !@DeNiro that acting was 🔥 🔥 😊 . I want sequel #Casino ! 😁



goodfellas authentic **kind** provide anchor genre pesci deniro liotta perfect popular **impression** casino way end acting **fire fire smile face**
smile eye want sequel casino **loudly cry face**

TF-IDF

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}} * IDF = \log\left(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}}\right)$$

- **Remove** all words that do not appear in at least 2 tweets or more than 90% of tweets.
- **Include** unigrams and bigrams

	goodfellas	authentic	kind	impression	face	goodfellas authentic	authentic kind
tweet1	0.2	0.03	0.1	0.08	0.9	0.1	0.5
tweet2	0	0	0.3	0.1	0.2	0	0
.....
.....
.....
tweet7204	0	0.4	0.8	0.8	0.1	0.01	0.07

Train set: n_samples: 7204, n_features: 16619

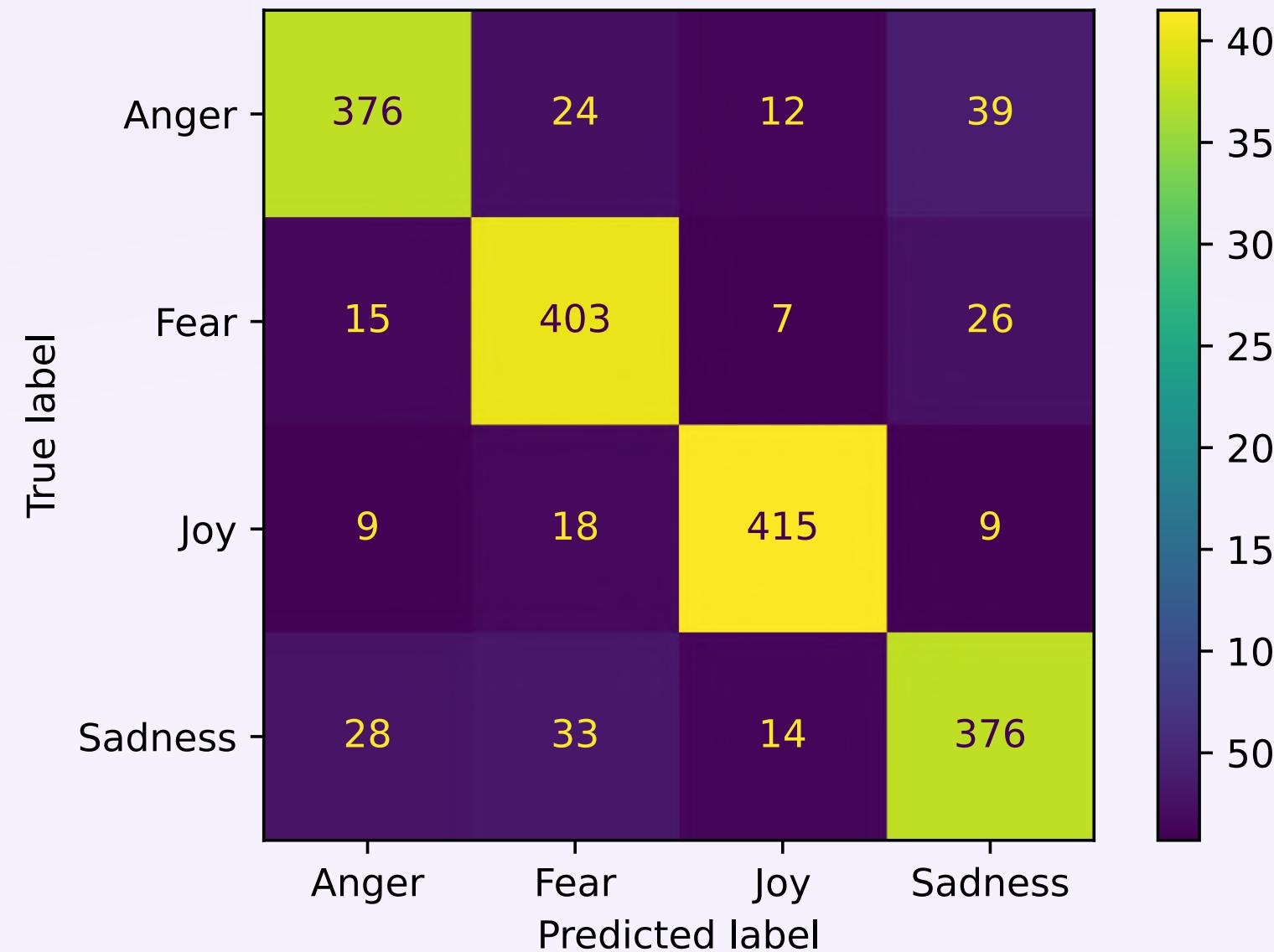
Test set: n_samples: 1804, n_features: 16619

Models

Algorithm\Score	F1 - train	F1 - test
SVC(SVM)	0.90	0.85
Passive aggressive	0.91	0.86
Random forest	0.88	0.84
SGD (hinge loss)*	0.91	0.87
Naive Bayes	0.87	0.82
Logistic regression	0.82	0.79

- Algorithms were tuned on the training set.
- Five-fold cross-validation was set.
- The best parameters were selected according to the highest averaged F1 score.

Models - confusion matrix



Cornell Movie-Dialogs Corpus*

- # of Speakers: 9035

- # of Utterances: 304713

- # of Conversations: 83097

- # of Movies: 617

meta.character_name		text	speaker	reply_to	conversation_id
	ACE	Now, instead of the cops only lookin' at Nicky...	u4331	L120769	L120769
	ACE	It turned out Phil Green, Mr Integrity, had a ...	u4331	None	L120737
	ACE	...was gonna stop what came up next at the cas...	u4331	None	L120733
	ACE	It's like everything else in this place. You d...	u4331	L120543	L120540
	ACE	Look at that. Look at this. There's nothin'.....	u4331	L120541	L120540

	MARINO	Get up.	u4358	L120206	L120206
	MARINO	Smarten up.	u4358	None	L119926
	MARINO	Of course.	u4358	L119905	L119904
	MARINO	Whoa, whoa, whoa.	u4358	None	L119858
	MARINO	Well, how come I laid nine?	u4358	None	L119662

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

The screenshot shows a Quora search result for the question "Is the movie Casino a sequel to Goodfellas?". The page includes the URL "https://www.quora.com › Is-the-movie-Casino-a-sequel-to...", the Quora logo, and navigation links for "About featured snippets" and "Feedback". Below the question, there is a "People also ask" section with three collapsed dropdowns: "What is the difference between Goodfellas and Casino?", "Is Casino similar to Goodfellas?", and "Why are Casino and Goodfellas so similar?".

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

[https://www.quora.com/Is-the-movie-Casino-a-sequel-to... ::](https://www.quora.com/Is-the-movie-Casino-a-sequel-to-Goodfellas)

Is the movie Casino a sequel to Goodfellas? - Quora

About featured snippets · Feedback

People also ask ::

What is the difference between Goodfellas and Casino? ▾

Is Casino similar to Goodfellas? ▾

Why are Casino and Goodfellas so similar? ▾

[https://www.reddit.com/r/movies/comments/8qjwvz/serious... ::](https://www.reddit.com/r/movies/comments/8qjwvz/serious/)

[serious] Casino vs. Goodfellas. Which is the better gangster ...

Mar 23, 2018 — In my opinion, **Casino** is the better gangster film. Bigger money, more violent murders, and it goes deeper into the gangster lifestyle, showing the home life as ...

Goodfellas or Casino: which is the better movie? - Reddit Jan 16, 2017

What movie completes the **Goodfellas and Casino** Trilogy? Jul 11, 2020

Does anyone here like **Casino** more than **Goodfellas** ... - Reddit Jan 6, 2020

Goodfellas or Casino: Which is the better film? - Reddit Dec 15, 2021

More results from www.reddit.com

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

[https://www.quora.com/Is-the-movie-Casino-a-sequel-to... ::](https://www.quora.com/Is-the-movie-Casino-a-sequel-to-Goodfellas)

Is the movie Casino a sequel to Goodfellas? - Quora

About featured snippets

People also ask ::

- What is the difference between Goodfellas and Casino?
- Is Casino similar to Goodfellas?
- Why are Casino and Goodfellas so similar?



vies > comments > serious... ::

Goodfellas. Which is the better gangster ...

on, Casino is the better gangster film. Bigger money, more violent into the gangster lifestyle, showing the home life as ...

h is the better movie? - Reddit Jan 16, 2017

Goodfellas and Casino Trilogy? Jul 11, 2020

no more than Goodfellas ... - Reddit Jan 6, 2020

h is the better film? - Reddit Dec 15, 2021

it.com

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

<https://www.quora.com/Is-the-movie-Casino-a-sequel-to-Goodfellas>

Is the movie Casino a sequel to Goodfellas? - Quora

About featured snippets

People also ask :

- What is the difference between Goodfellas and Casino?
- Is Casino similar to Goodfellas?
- Why are Casino and Goodfellas so similar?



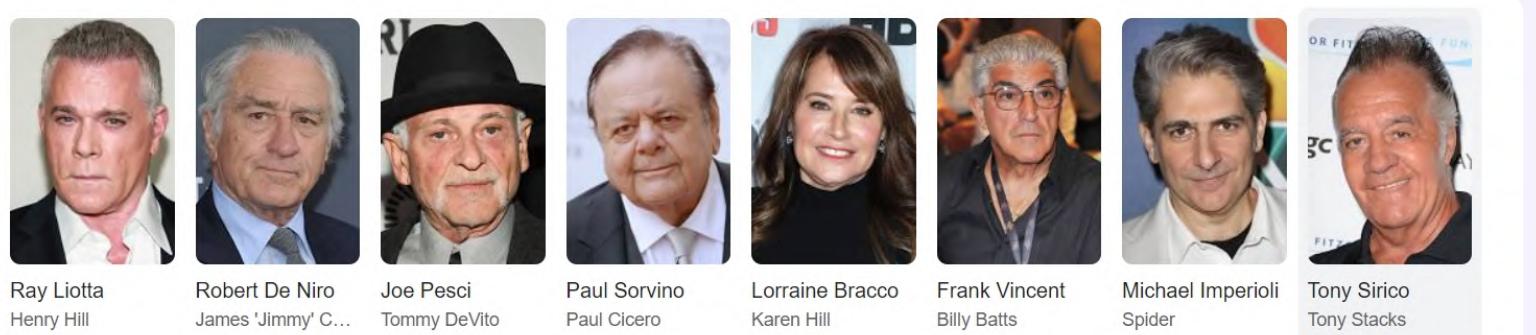
Robert De Niro Sam 'Ace' Rothstein
Joe Pesci Nicky Santoro
Sharon Stone Ginger McKenna
Frank Vincent Frank Marino
John Manca Wiseguy Eddy
James Woods Lester Diamond
Don Rickles Billy Sherbert
Cathy Scorsese Piscano's Daughter

vies › comments › serious... ::

Goodfellas. Which is the better gangster ...

on, **Casino** is the better gangster film. Bigger money, more violent into the gangster lifestyle, showing the home life as ...

h is the better movie? - Reddit Jan 16, 2017
Goodfellas and Casino Trilogy? Jul 11, 2020
no more than Goodfellas ... - Reddit Jan 6, 2020
h is the better film? - Reddit Dec 15, 2021
www.reddit.com



Ray Liotta Henry Hill
Robert De Niro James 'Jimmy' C...
Joe Pesci Tommy DeVito
Paul Sorvino Paul Cicero
Lorraine Bracco Karen Hill
Frank Vincent Billy Batts
Michael Imperioli Spider
Tony Sirico Tony Stacks

Casino and Goodfellas emotions

Best model is used to predict emotions in utterances of "Casino" and "Goodfellas" movies.

But **why** exactly those two movies?

https://www.quora.com › Is-the-movie-Casino-a-sequel-to... ::

Is the movie Casino a sequel to Goodfellas? - Quora

About featured snippets

People also ask :

- What is the difference between Goodfellas and Casino?
- Is Casino similar to Goodfellas?
- Why are Casino and Goodfellas so similar?

Similar way of writing!

Ray Liotta Henry Hill
Robert De Niro James 'Jimmy' C...
Joe Pesci Tommy DeVito
Paul Sorvino Paul Cicero
Lorraine Bracco Karen Hill
Frank Vincent Billy Batts
Michael Imperioli Spider
Tony Sirico Tony Stacks

Robert De Niro Sam 'Ace' Rothstein
Joe Pesci Nicky Santoro
Sharon Stone Ginger McKenna
Frank Vincent Frank Marino
John Manca Wiseguy Eddy
James Woods Lester Diamond
Don Rickles Billy Sherbert
Cathy Scorsese Piscano's Daughter

Casino characters

CASINO



ACE: Ambitious and focused, but **vulnerable** to his emotions, particularly when it comes to his relationships with women.



NICKY: Aggressive and impulsive, but also fiercely loyal to those he considers family.



GINGER: Trapped and desperate, but also cunning and **manipulative**.



LESTER: Energetic and confident, but also reckless and self-destructive.

Goodfellas characters

GoodFellas



HENRY: Torn between his loyalty to the mafia and his desire for a normal life, but ultimately **unable** to escape his criminal past.



JIMMY: **Ruthless** and **paranoid**, but also charming and charismatic.



TOMMY: **Impulsive** and **violent**, but also quick-witted and loyal to his friends.

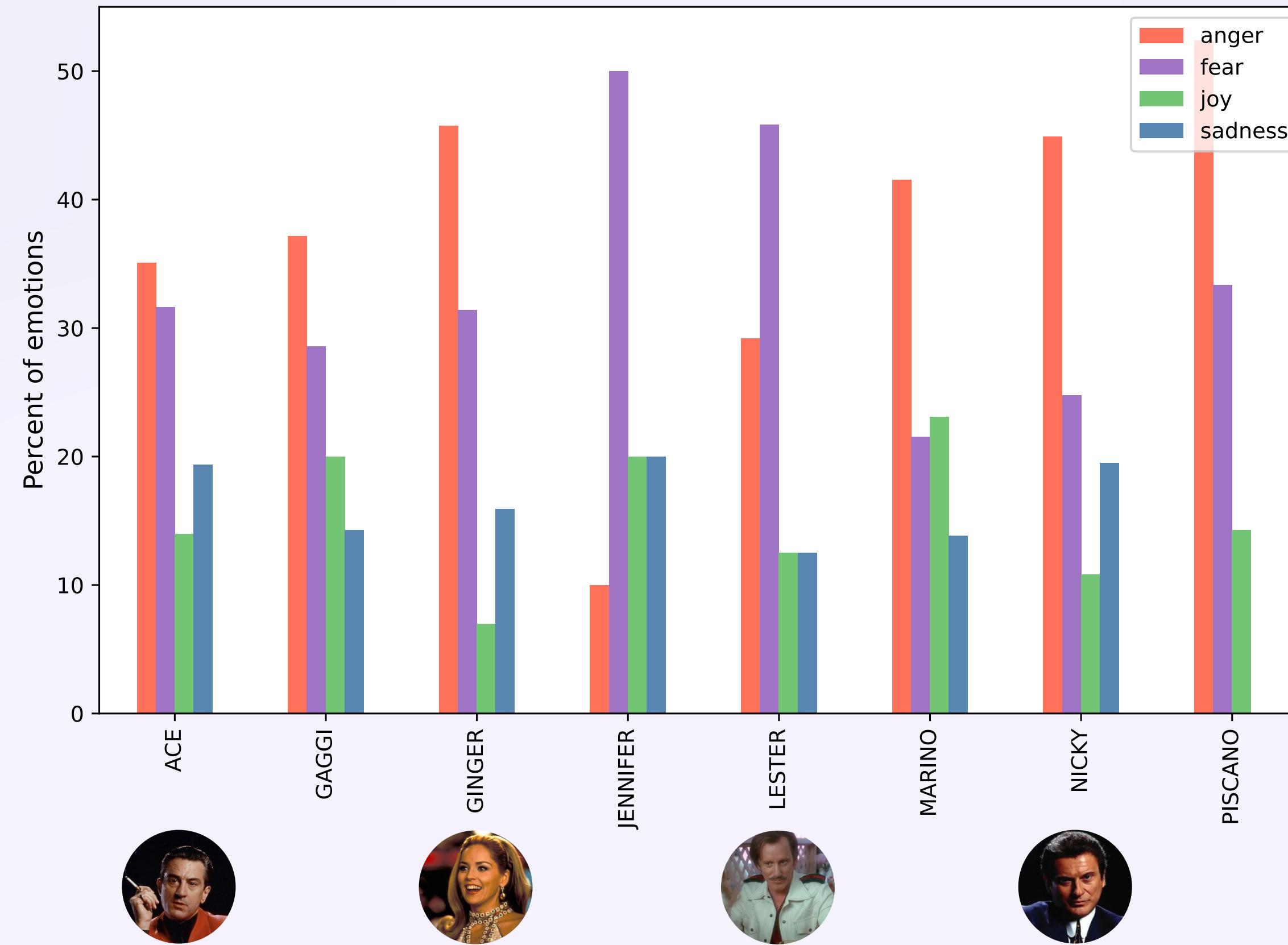


KAREN: Devoted and **loving**, but also consumed by anger and **jealousy** towards her husband's criminal lifestyle.

Casino characters - Emotions

CASINO

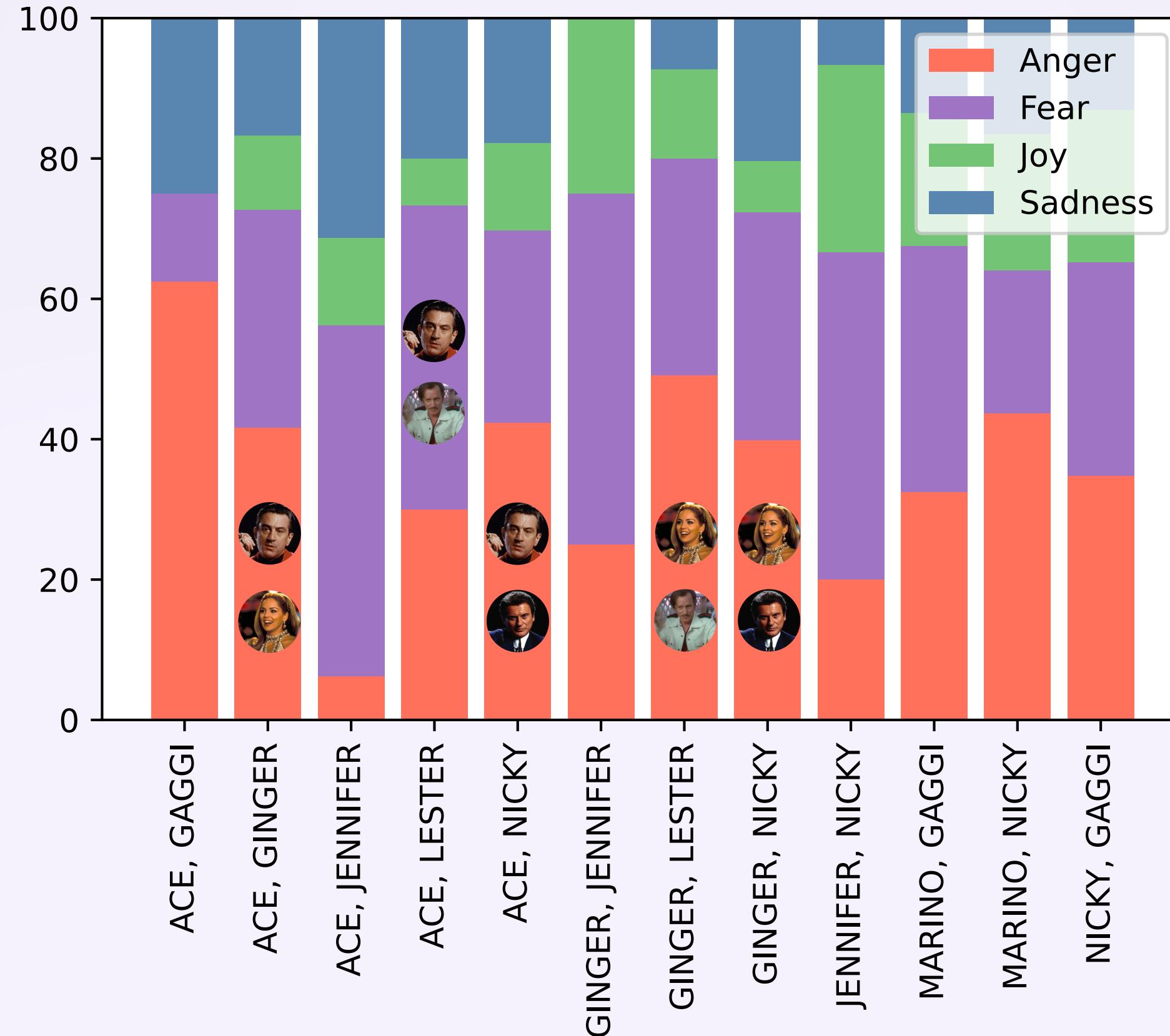
Characters emotions



Casino characters - Emotions

CASINO

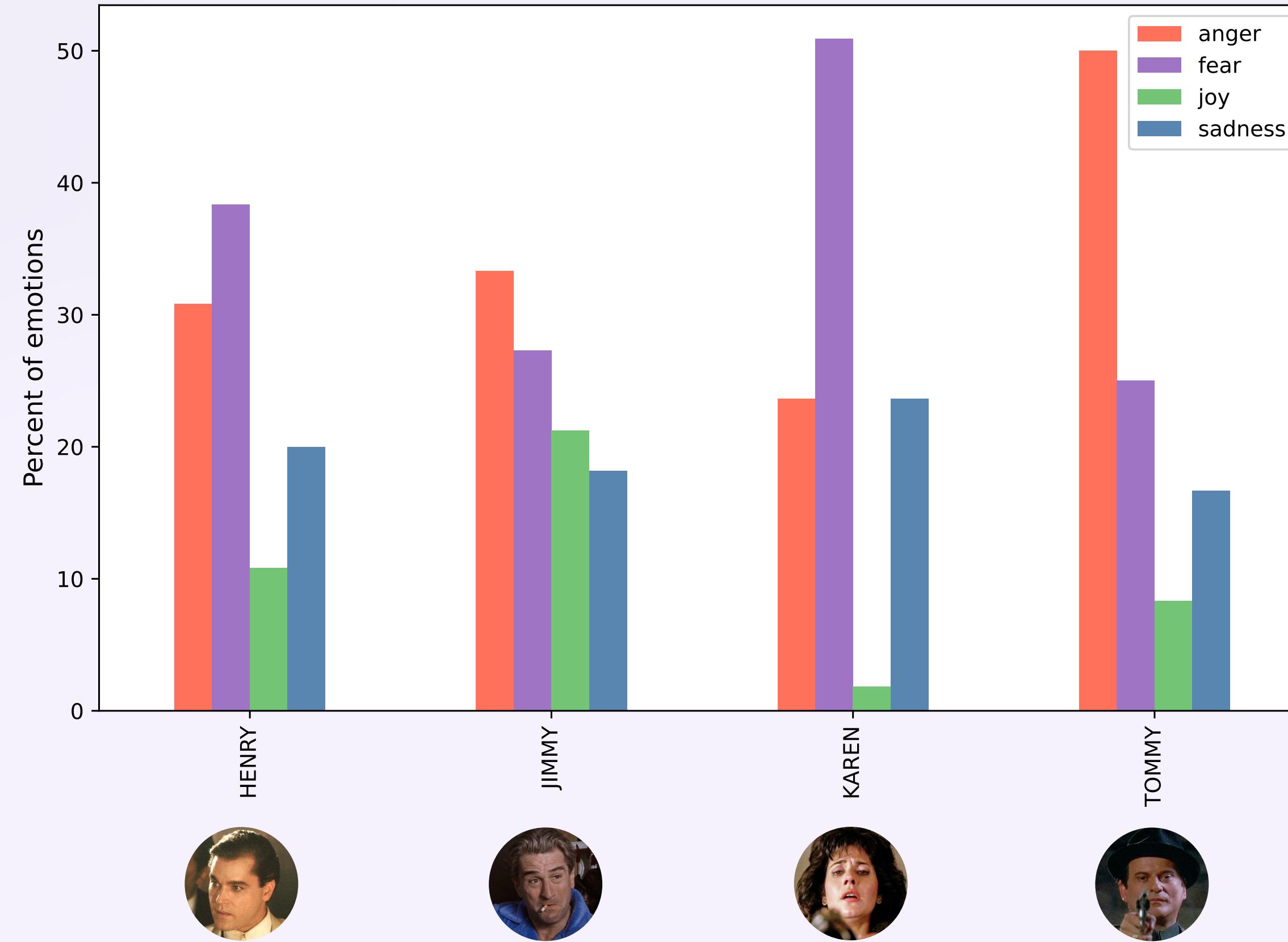
Character interactions



Goodfellas characters - Emotions

GoodFellas

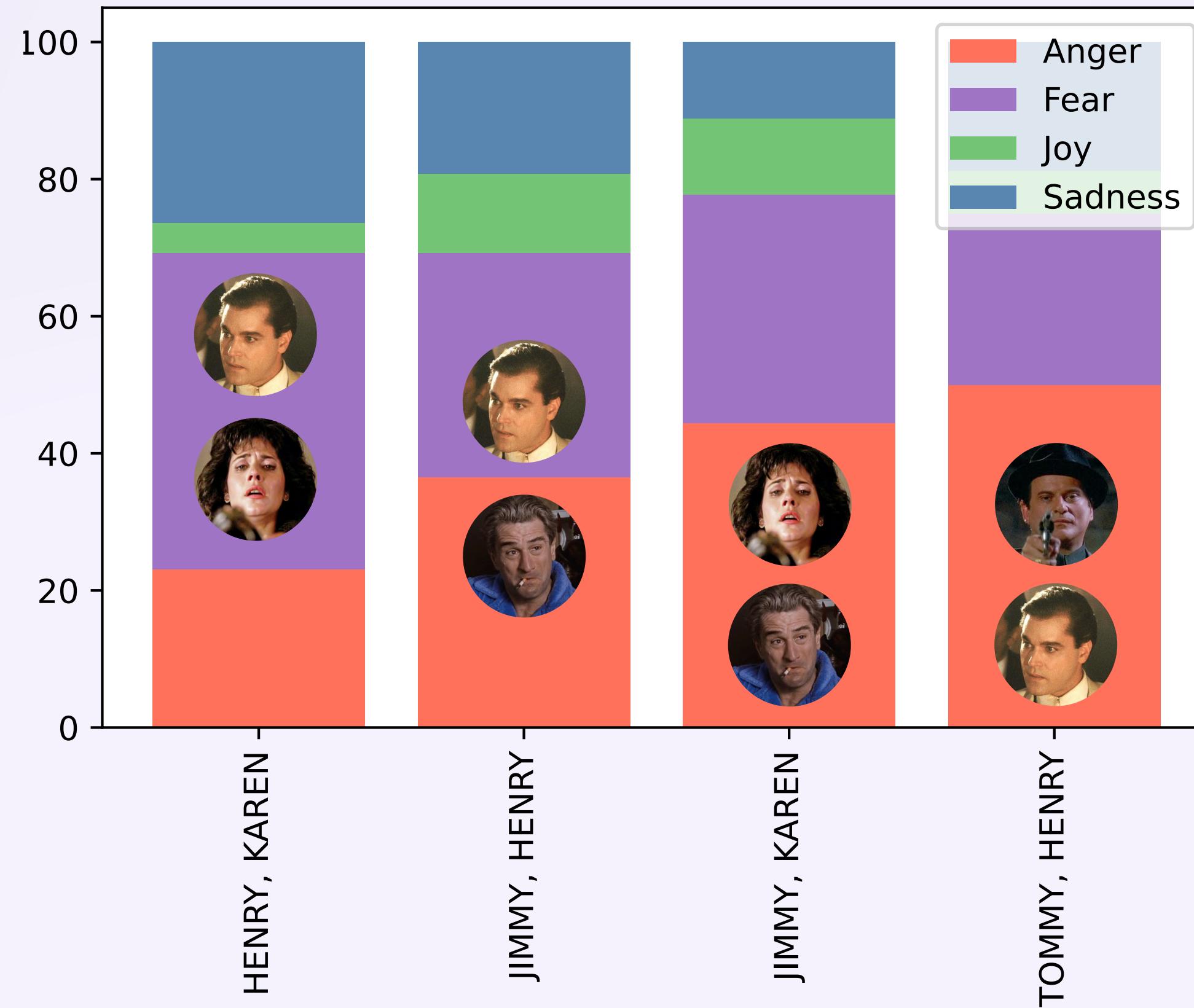
Characters emotions



Goodfellas characters - Emotions

GoodFellas

Character interactions



Subtitles files*

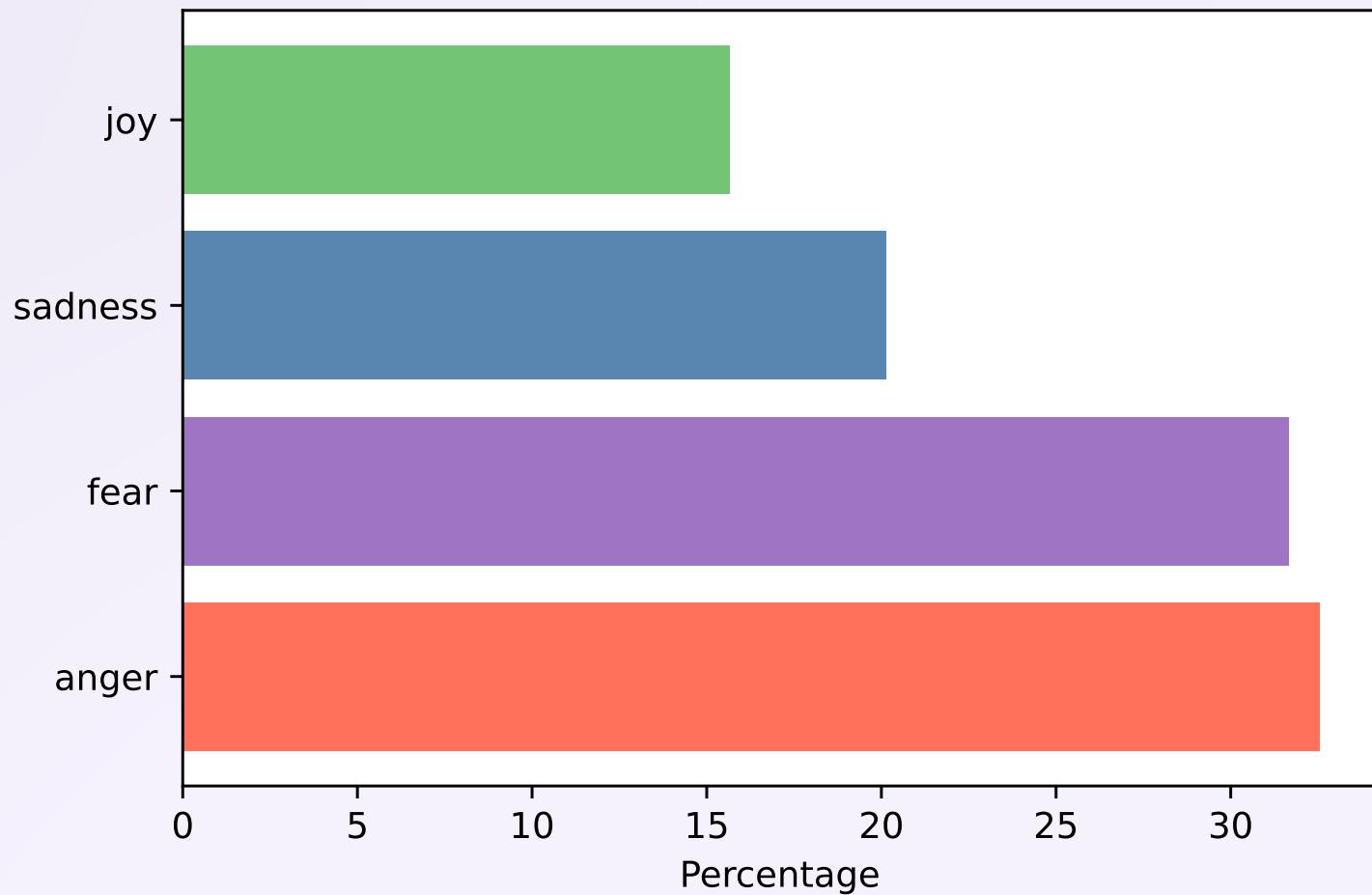
- The Cornell Movie dataset is helpful but it **does not have all utterances**.
- If we want to investigate emotions through the movie, there are **no timestamps**.
- A possible solution to understand the overall sentiment of the movies is to use **subtitle files**.

Subtitles files

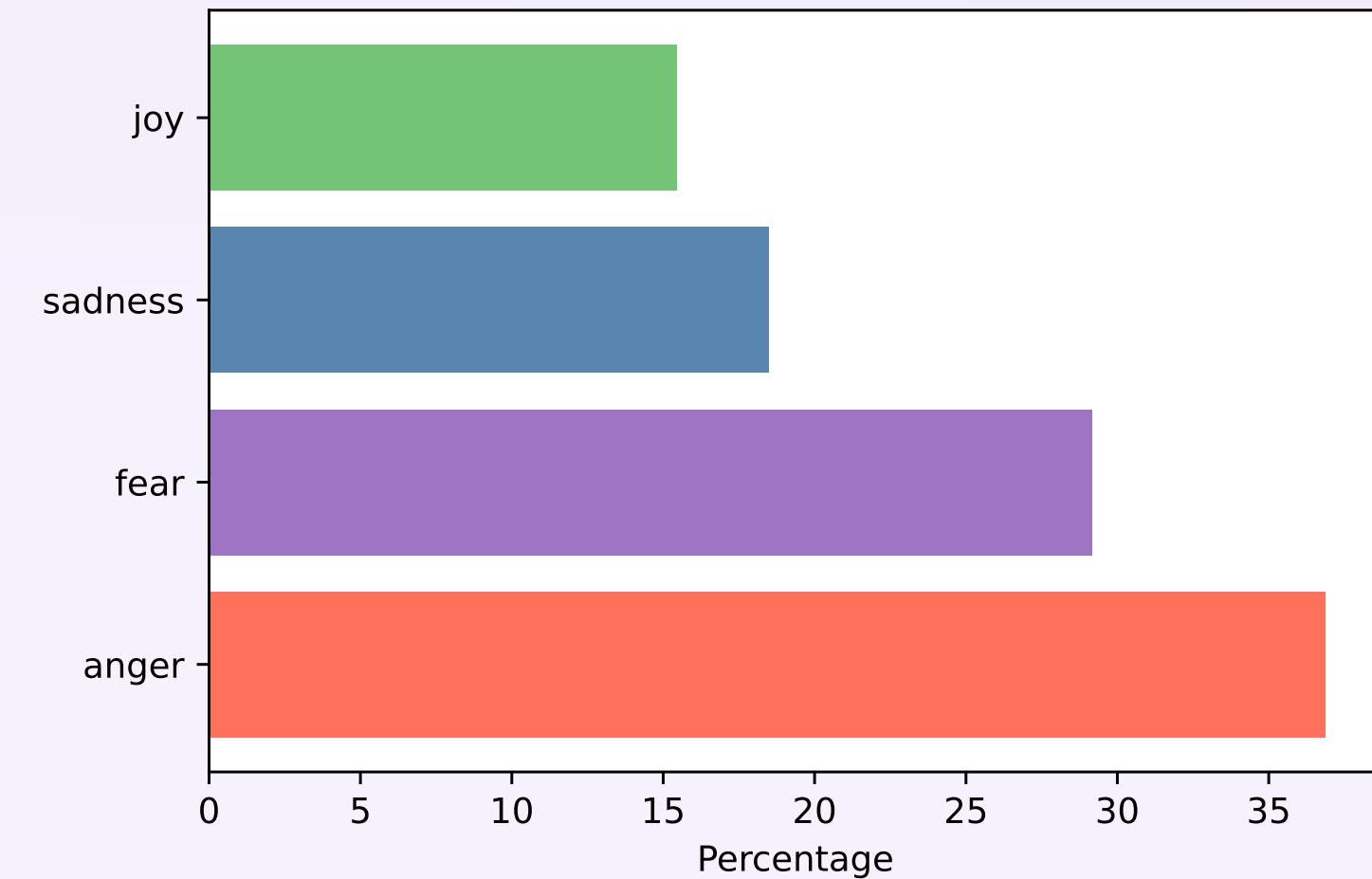
key	start_time	end_time	text
1	00:00:50,825	00:00:53,195	When you love someone, you've gotta trust them.
2	00:00:53,235	00:00:55,160	There's no other way.
3	00:00:55,195	00:00:58,874	You've got to give them the key to everything ...
4	00:00:59,374	00:01:03,164	Otherwise, what's the point?
5	00:01:03,664	00:01:08,333	And, for a while, I believed that's the kind o...
6	00:03:14,202	00:03:17,242	Before I ever ran a casino, or got myself blow...
7	00:03:17,332	00:03:21,502	Ace Rothstein was a hell of a handicapper, I c...
8	00:03:21,582	00:03:24,172	I was so good that whenever I bet, I could cha...
9	00:03:24,212	00:03:26,171	for every bookmaker in the country.
10	00:03:26,211	00:03:28,711	I'm serious, I had it down so cold...

Overall emotions

CASINO

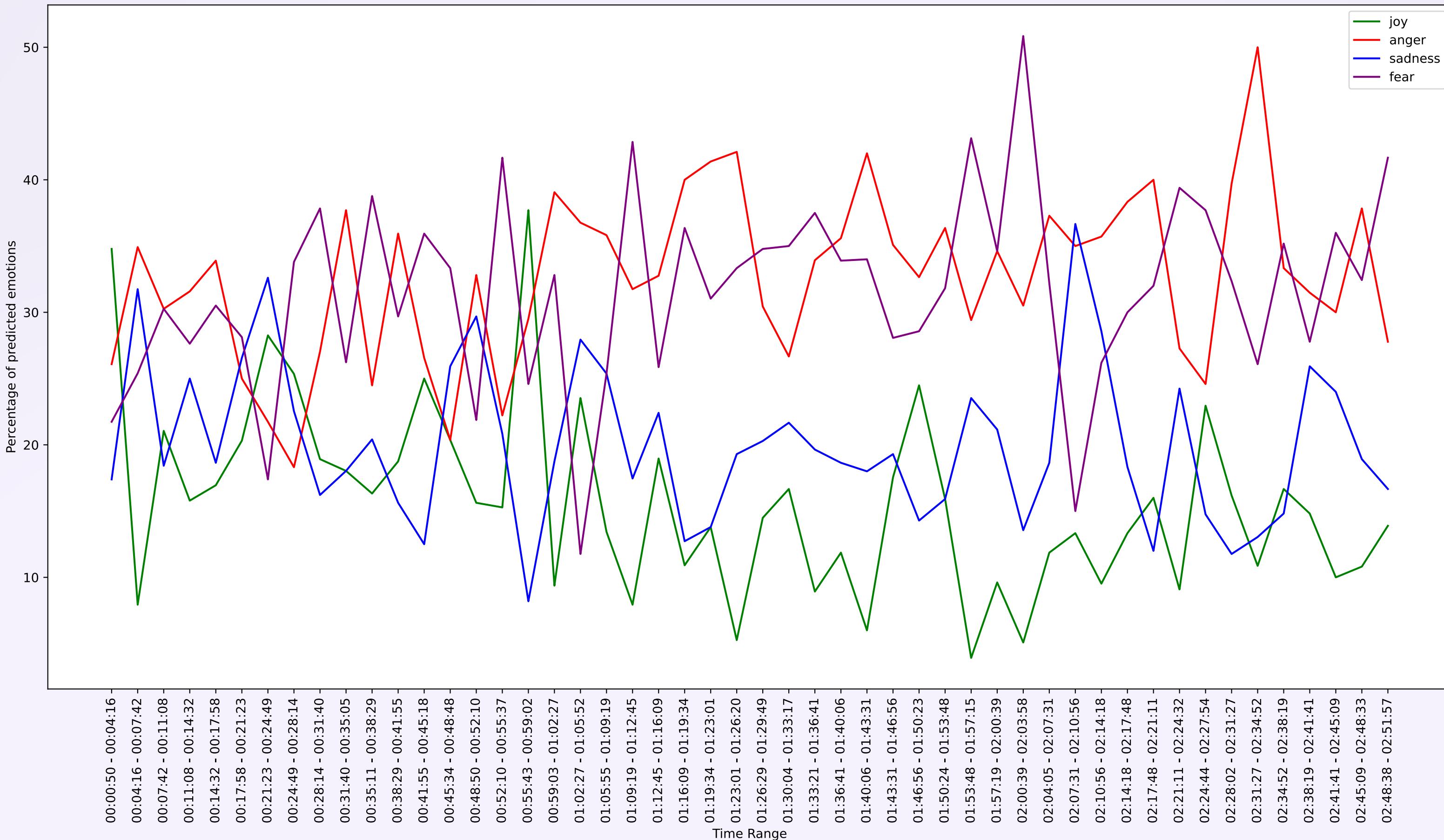


GoodFellas



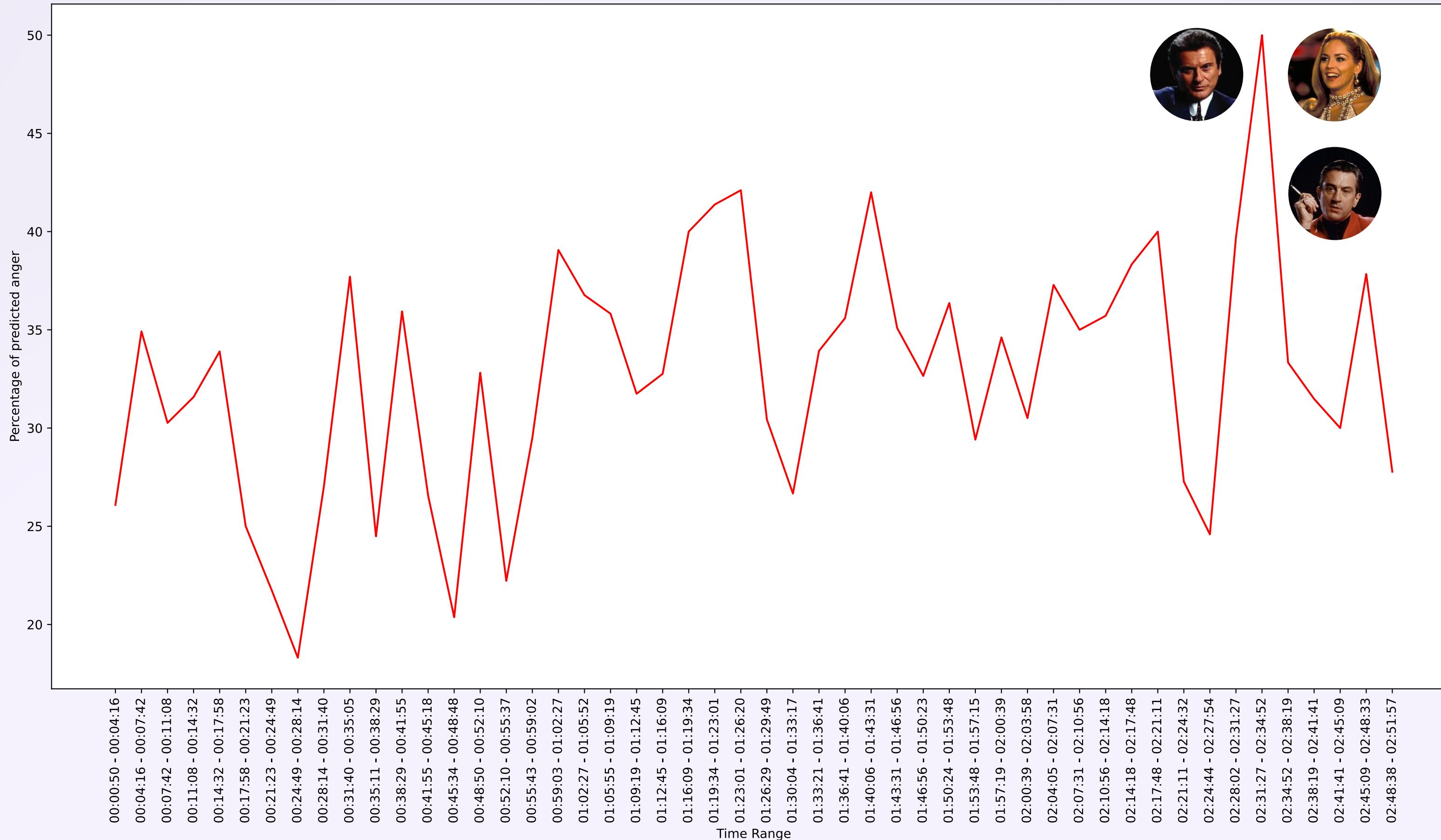
Overall emotions

CASINO



Peak of Anger

CASINO



Peak of Anger

CASINO



I should never have gotten involved
with this fu~~CENSORED~~ nut!



You'll ~~CENSORED~~ kill her.
Take it easy.



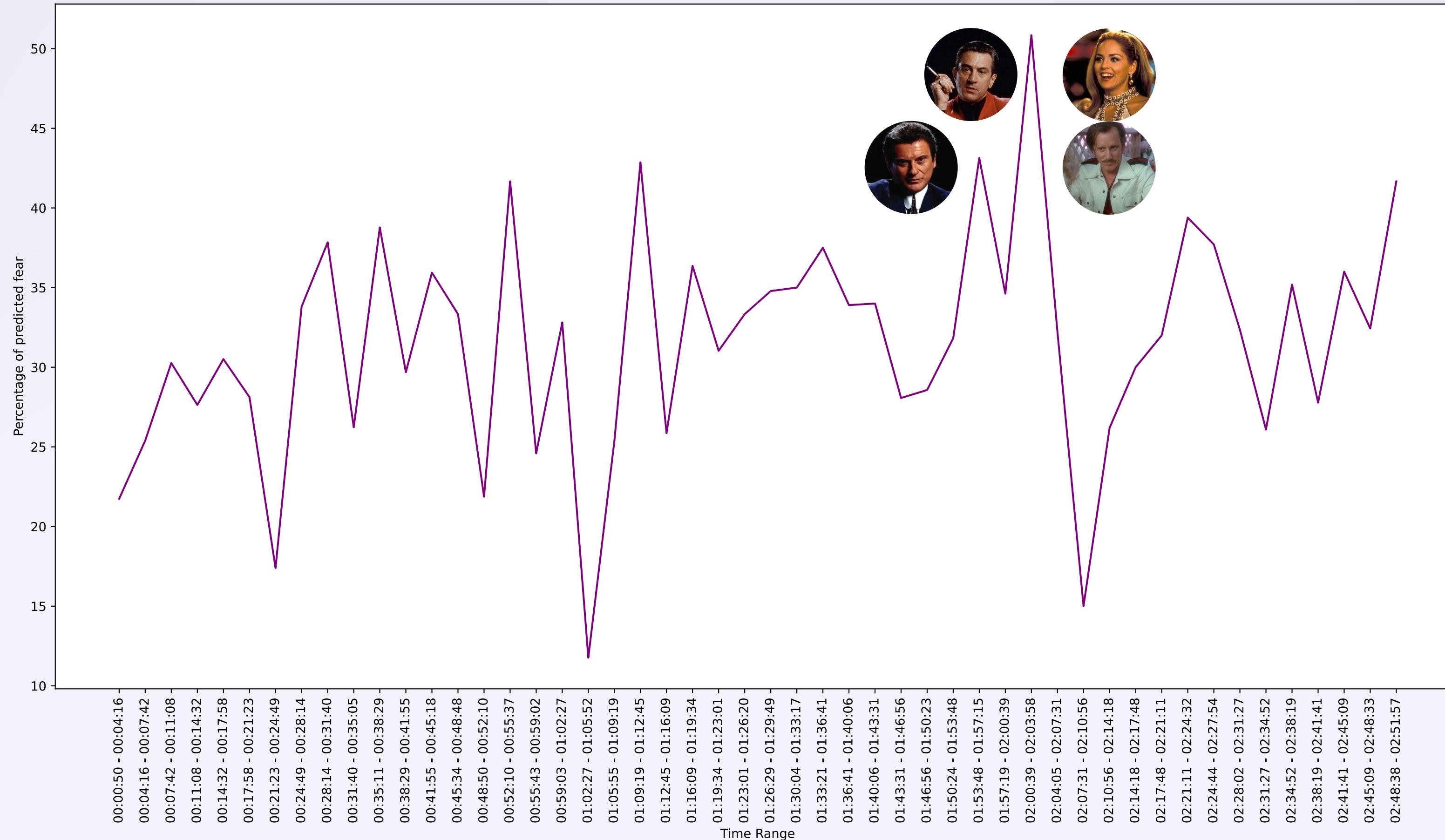
I don't need you!
I'll get my own fu~~CENSORED~~ money!



Get down here
and talk to me, goddamn it!

Peak of Fear

CASINO



Peak of Fear

CASINO

My daughter is with them, and I think they're gonna try and kidnap her.

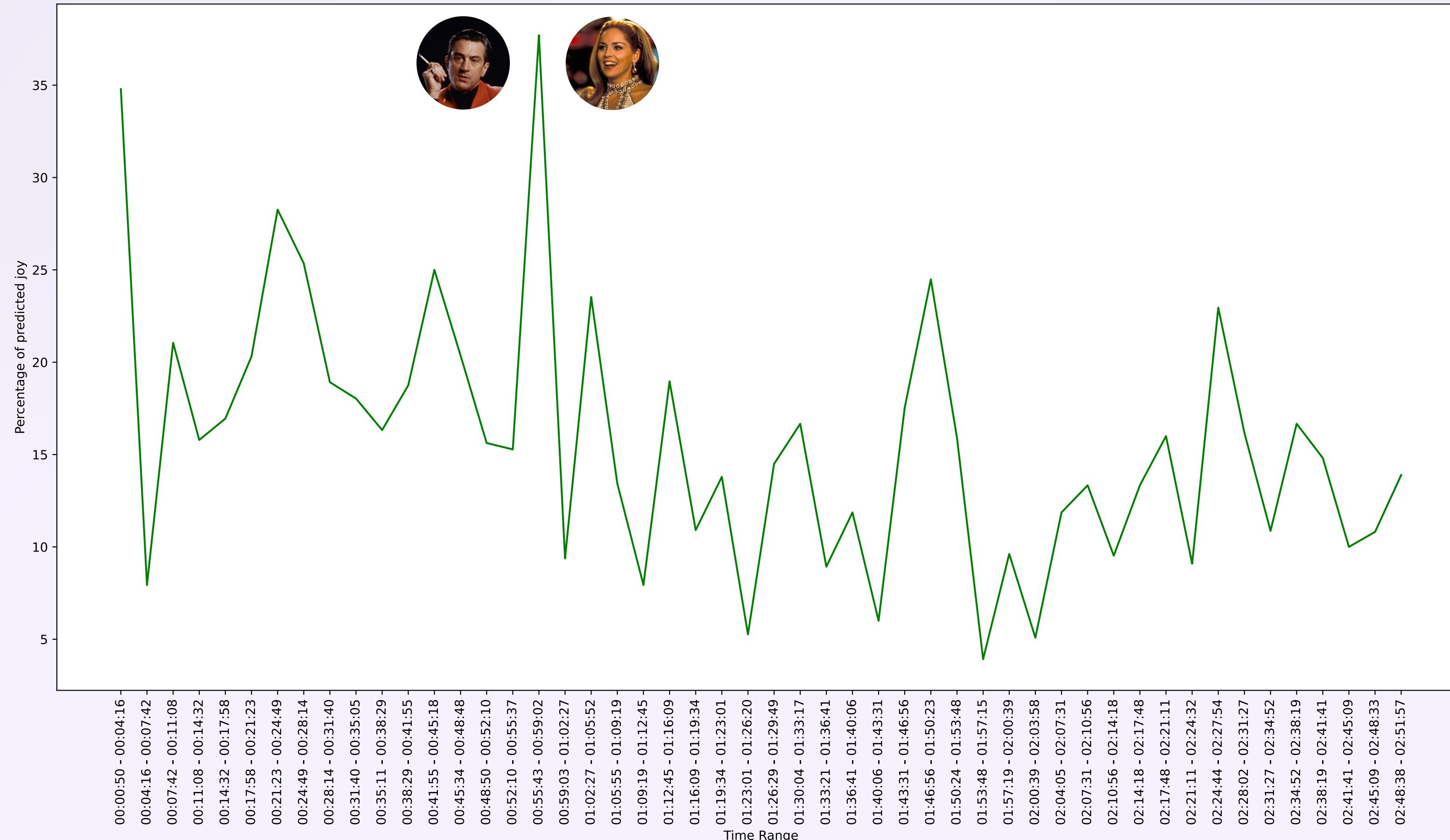
- I just talked to him.
- So he knows where you are.

What do you think we're gonna do? He's probably got guys outside the **CENSORED** house!

I think he's gonna kill me.
- Just relax.

Peak of Joy

CASINO



Try Pitch

Peak of Joy

CASINO



It is with great pleasure that I accept
this certificate of appreciation...



But my greatest pleasure was watching
my wife, Ginger, work the room.



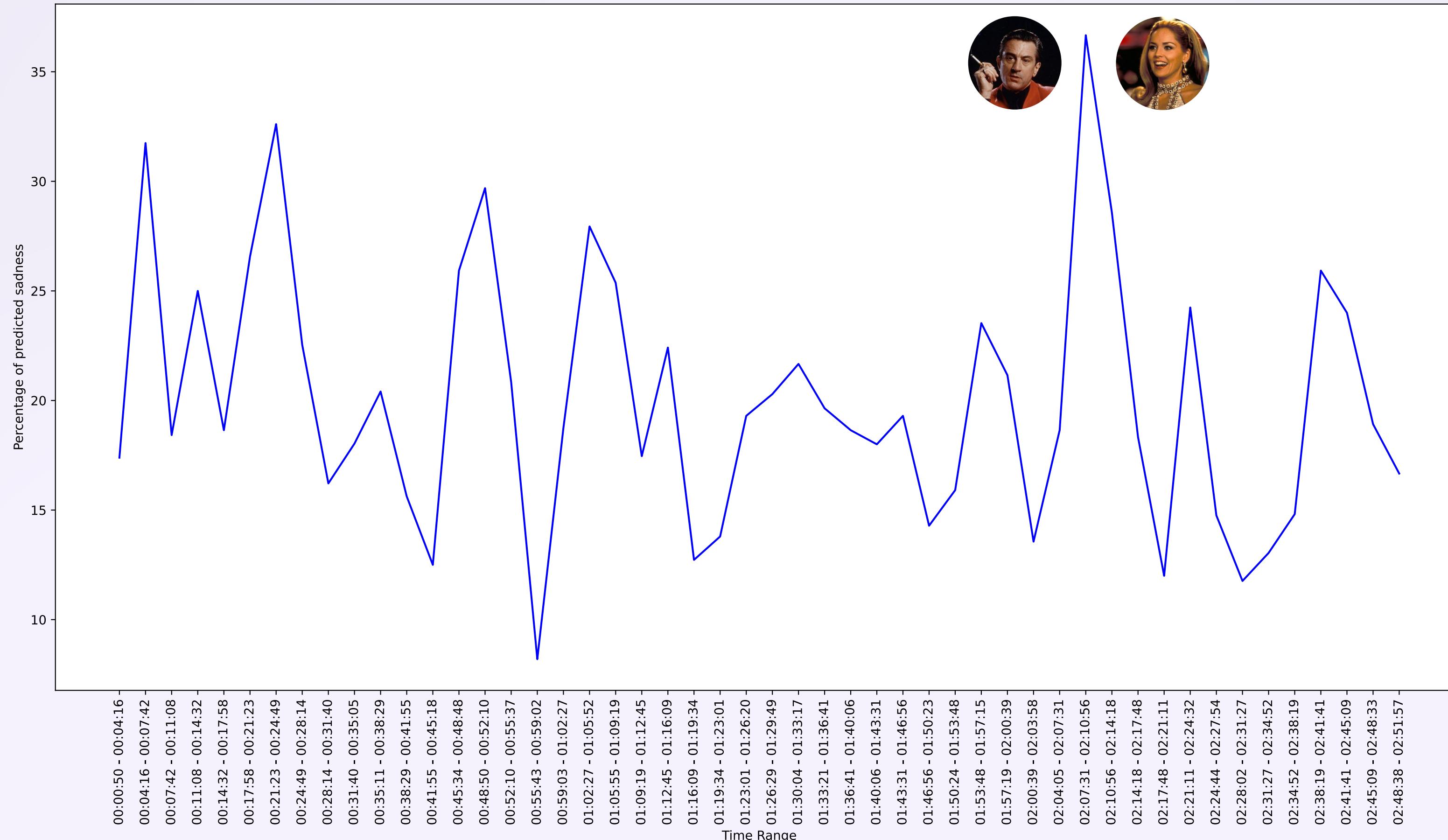
She made everybody feel good.



Thank you.
Thanks for that compliment.

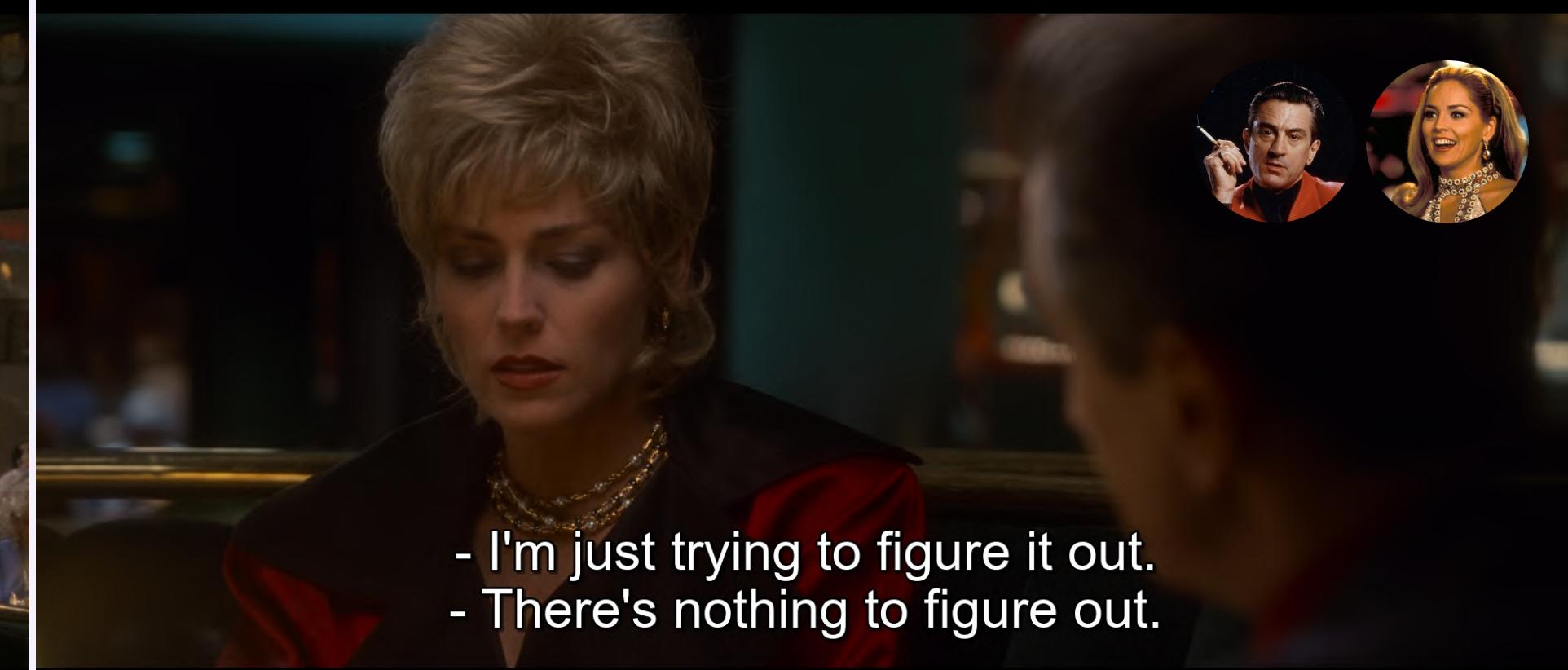
Peak of Sadness

CASINO



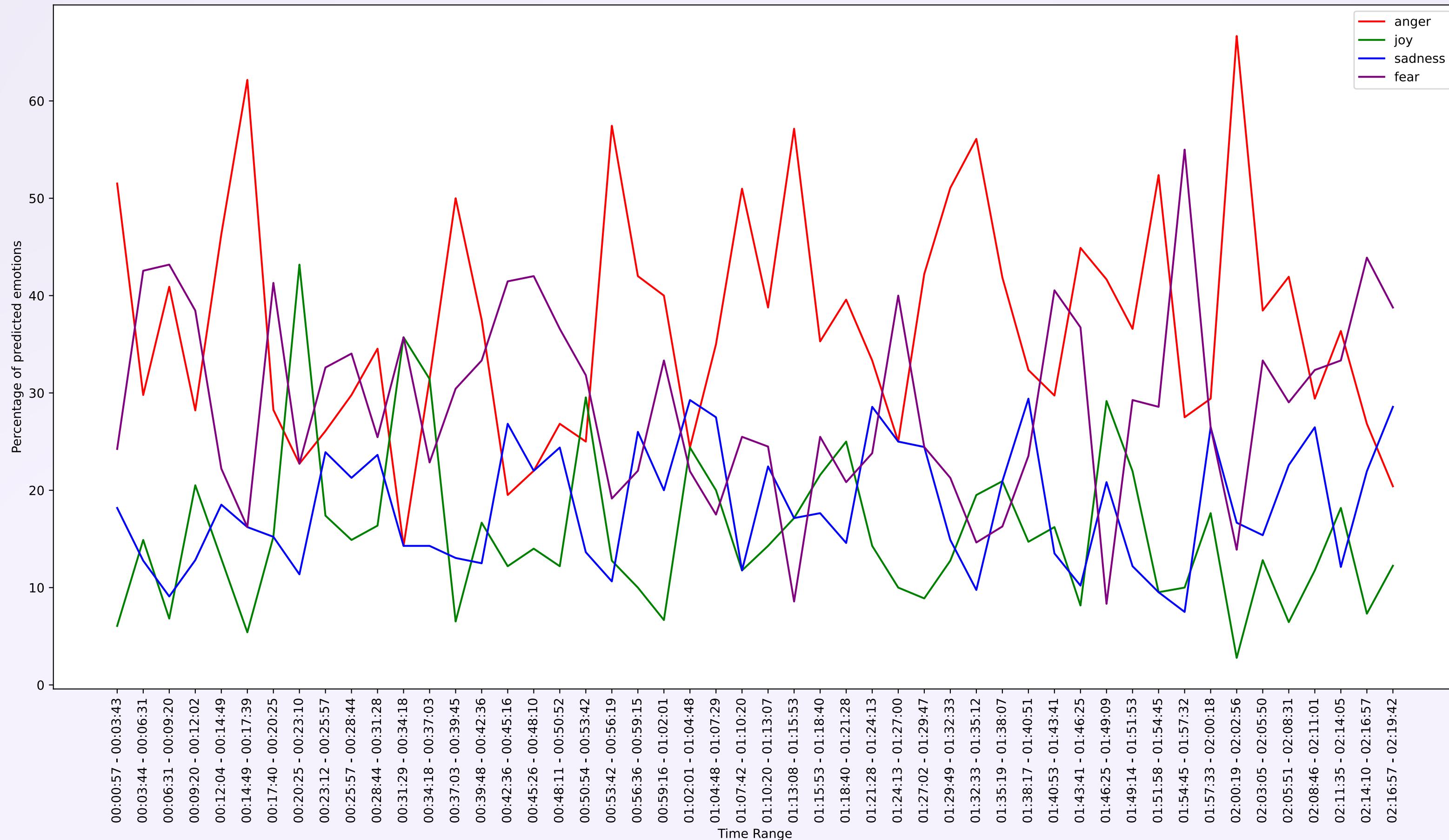
Peak of Sadness

CASINO



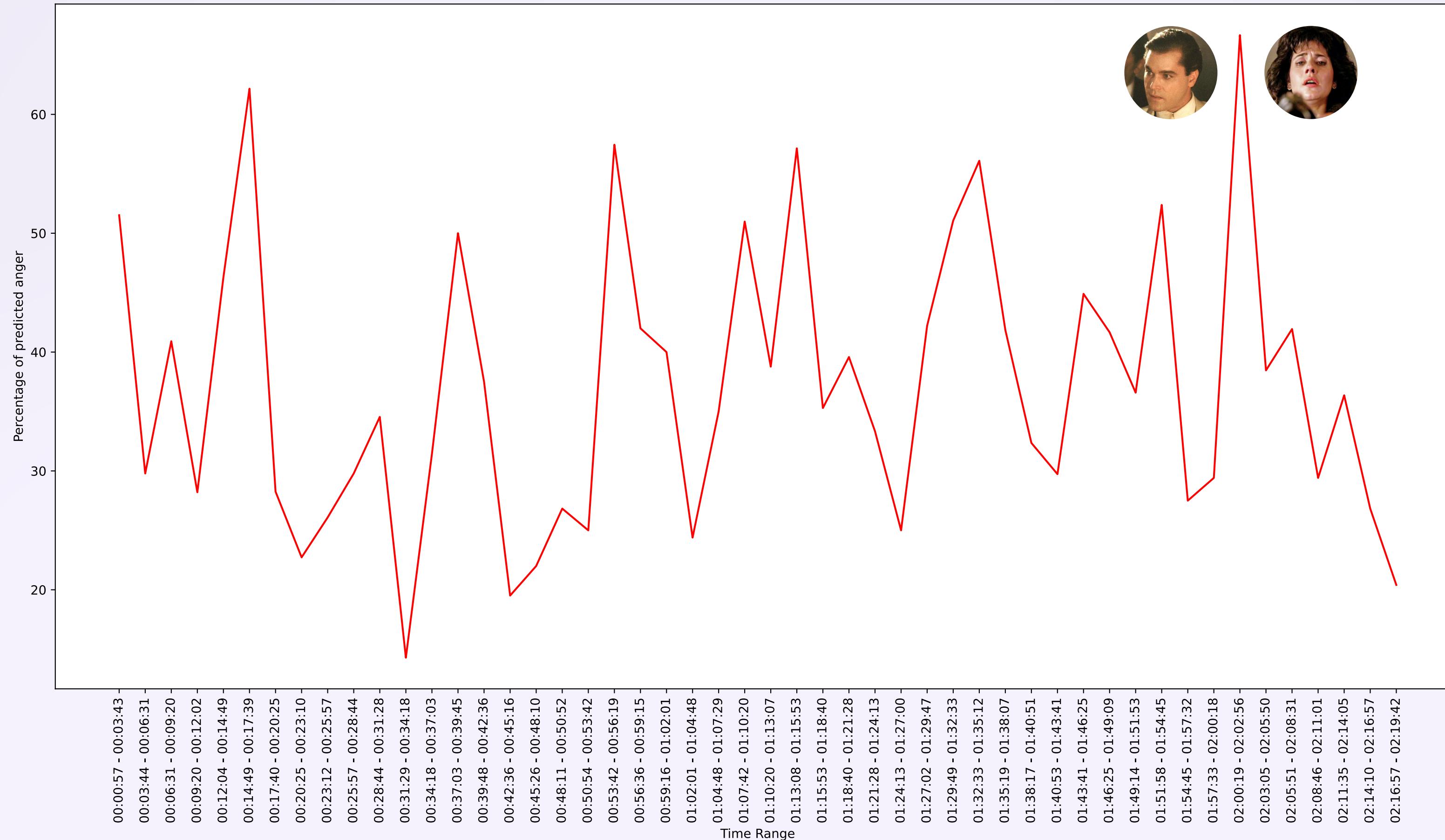
Overall emotions

GoodFellas



Peak of Anger

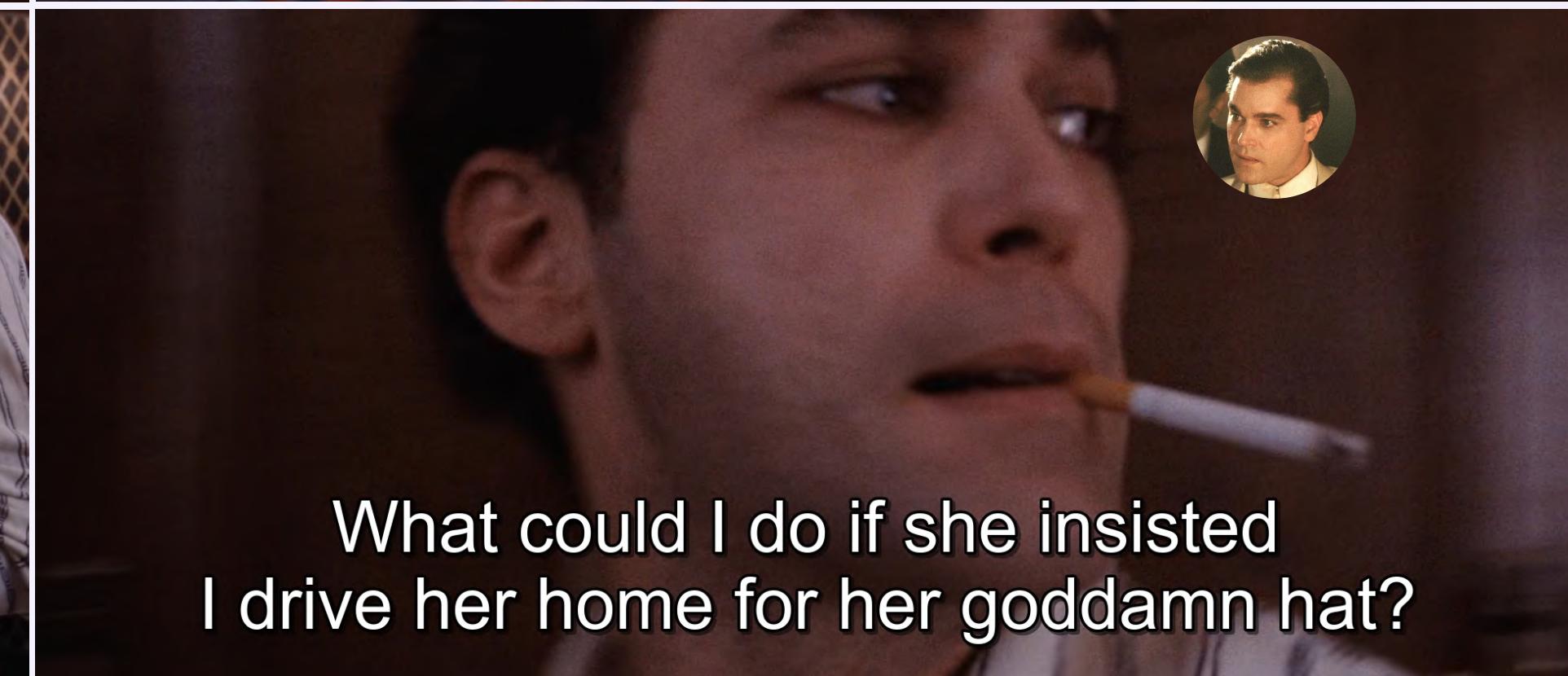
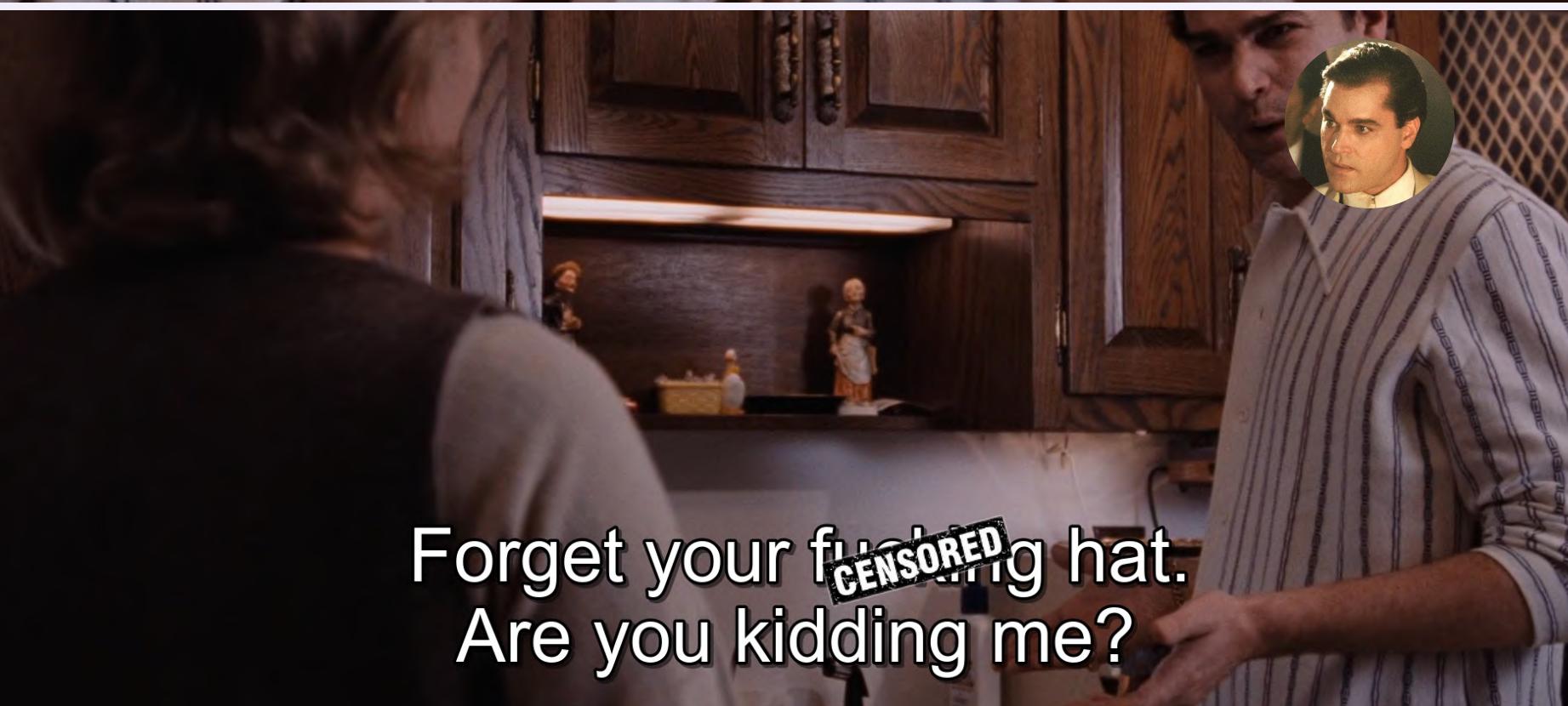
GoodFellas



Try Pitch

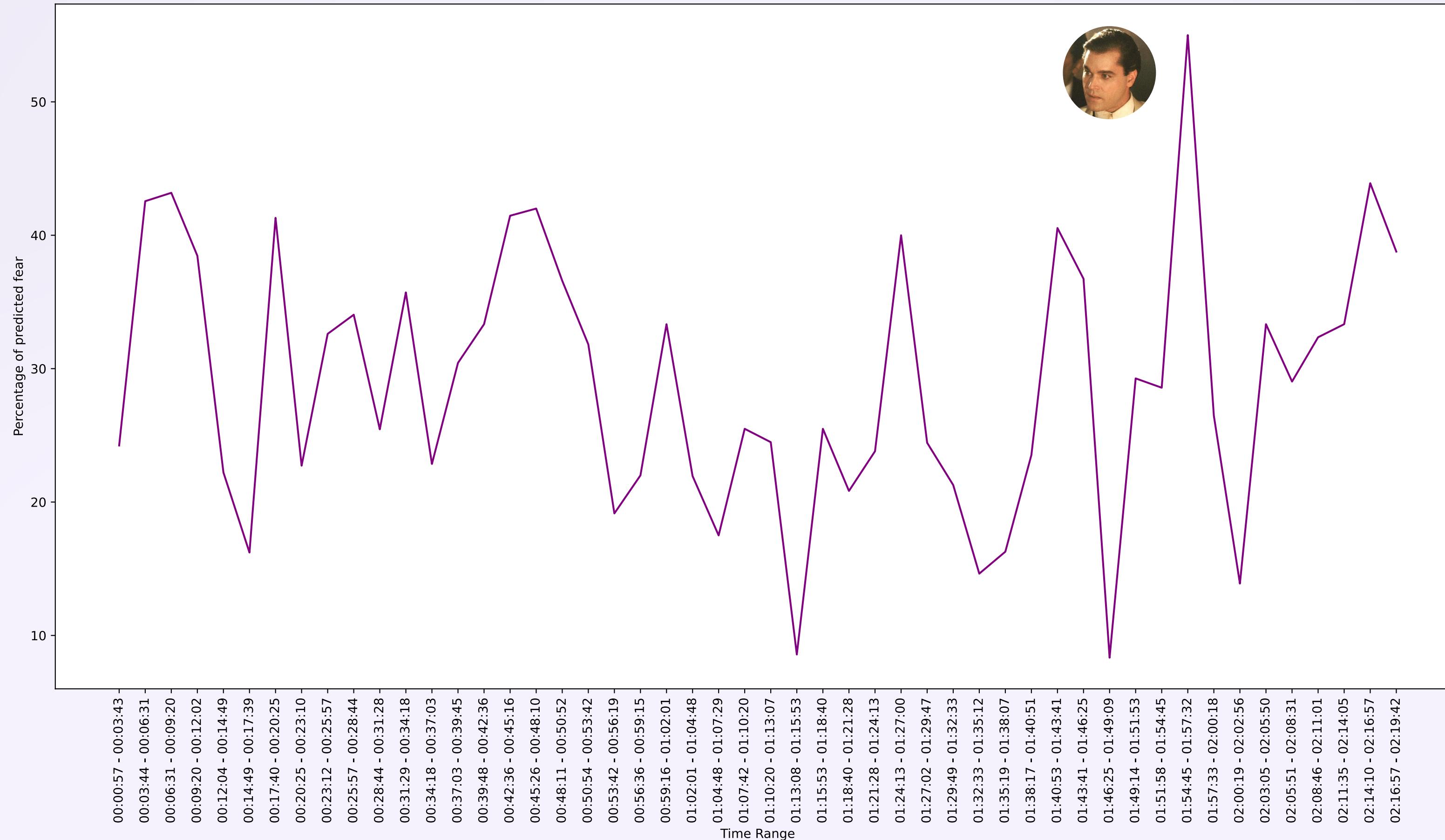
Peak of Anger

GoodFellas



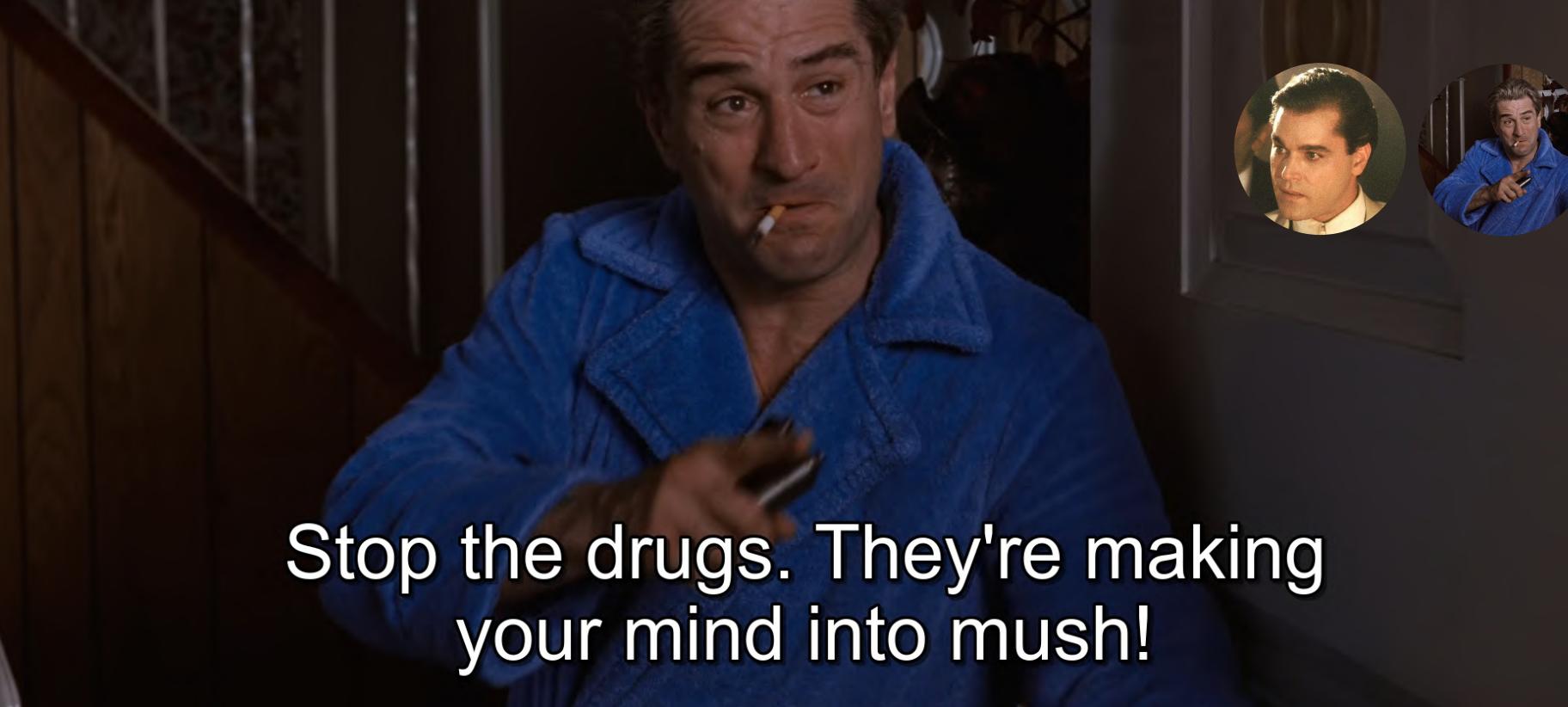
Peak of Fear

GoodFellas



Try Pitch

Peak of Fear



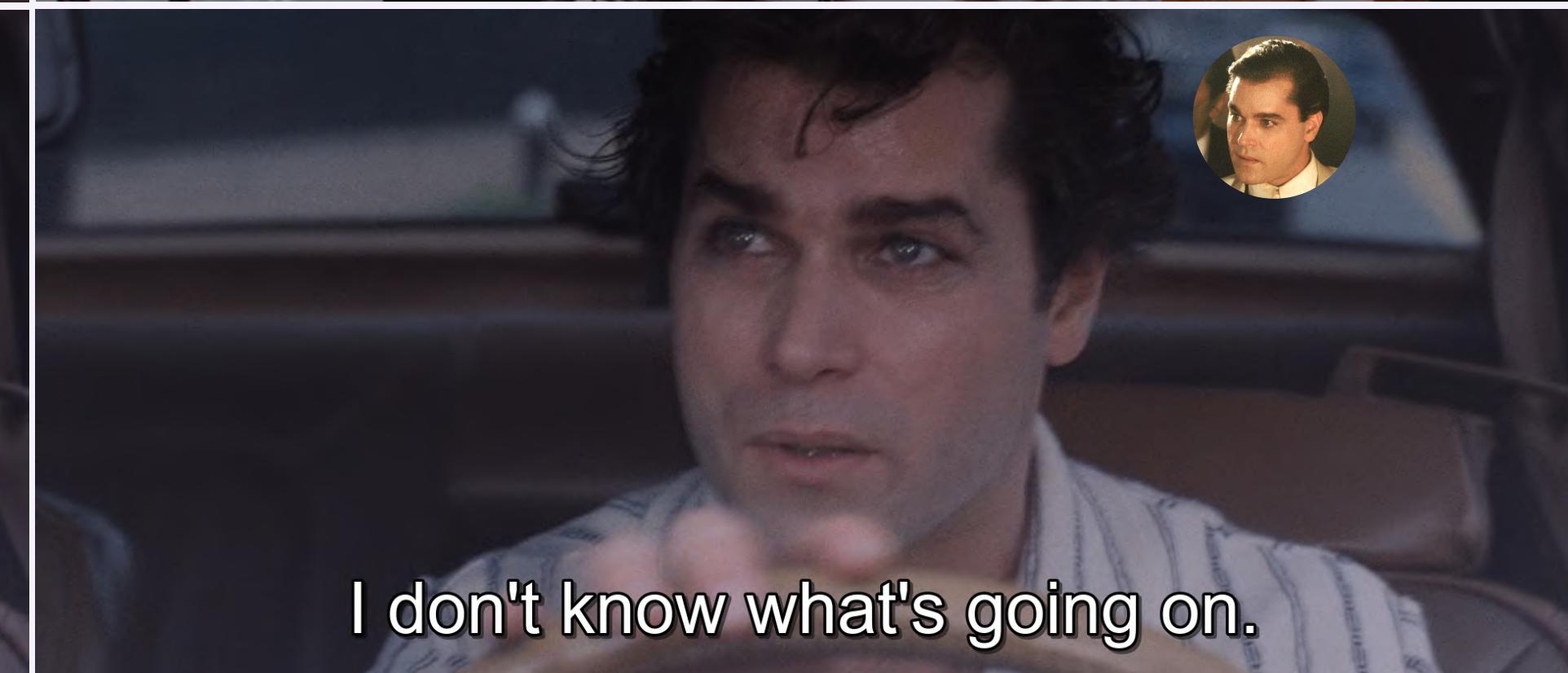
Stop the drugs. They're making
your mind into mush!



Jesus Christ, what ha,
~~CENSORED~~ppened to you?



Right in front ~~of~~, there.
I think it's been following me all morning.

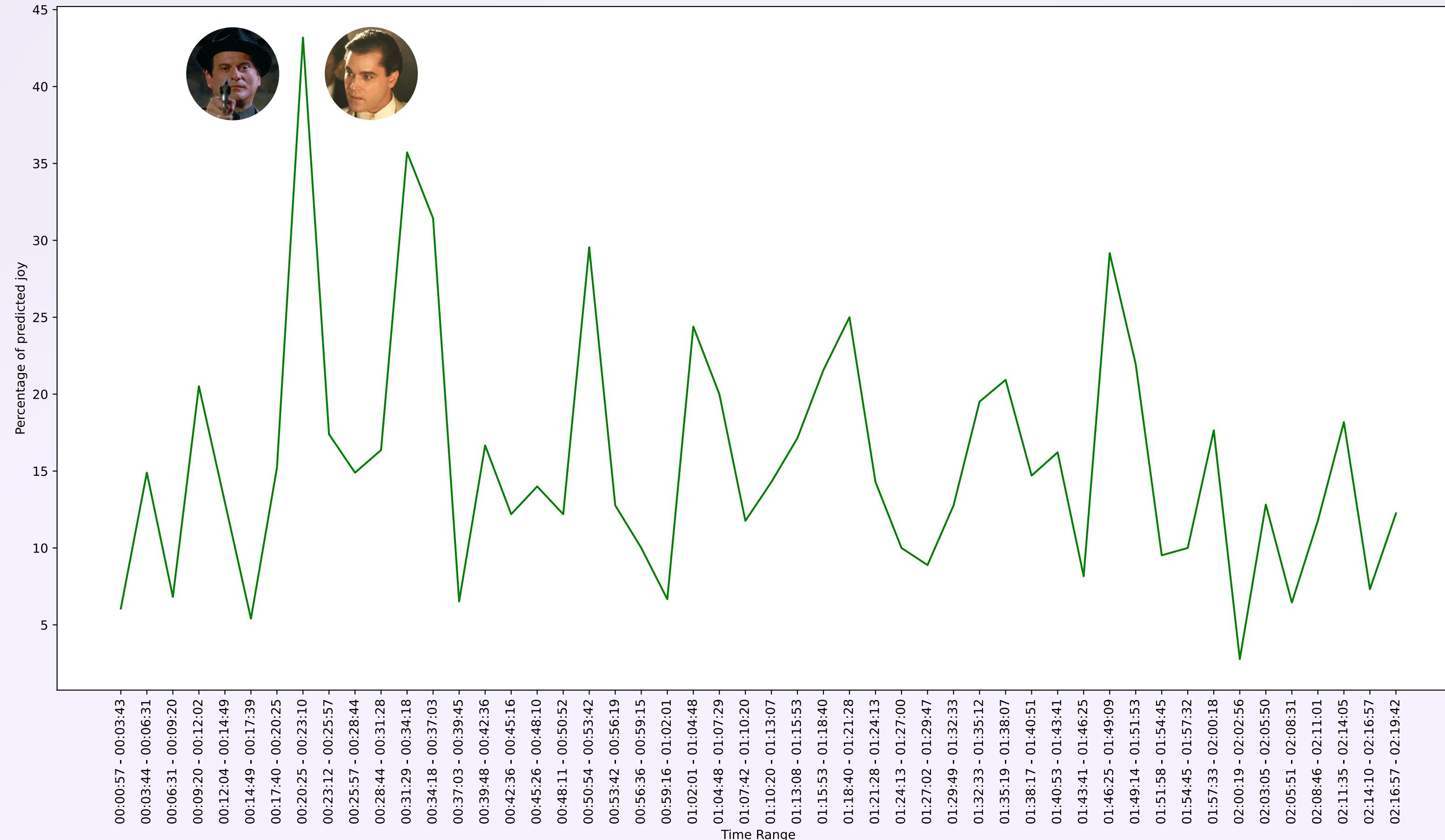


I don't know what's going on.

GoodFellas

Peak of Joy

GoodFellas



Peak of Joy

GoodFellas



Pow! Bing! Boom!



Funny. You're really funny.



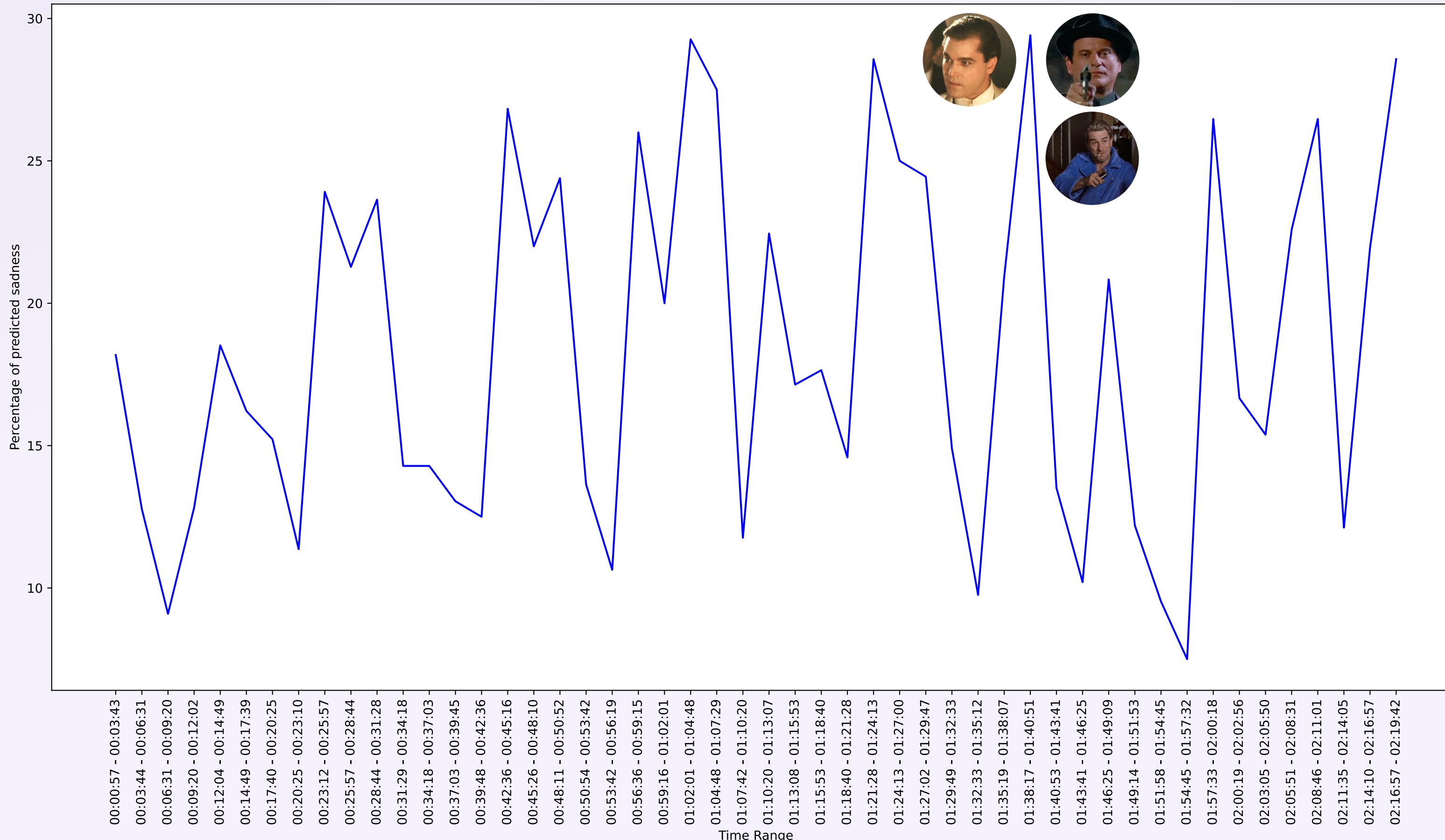
It's funny, you know. It's a good story.



What do you mean, funny?
How am I funny?

Peak of Sadness

GoodFellas



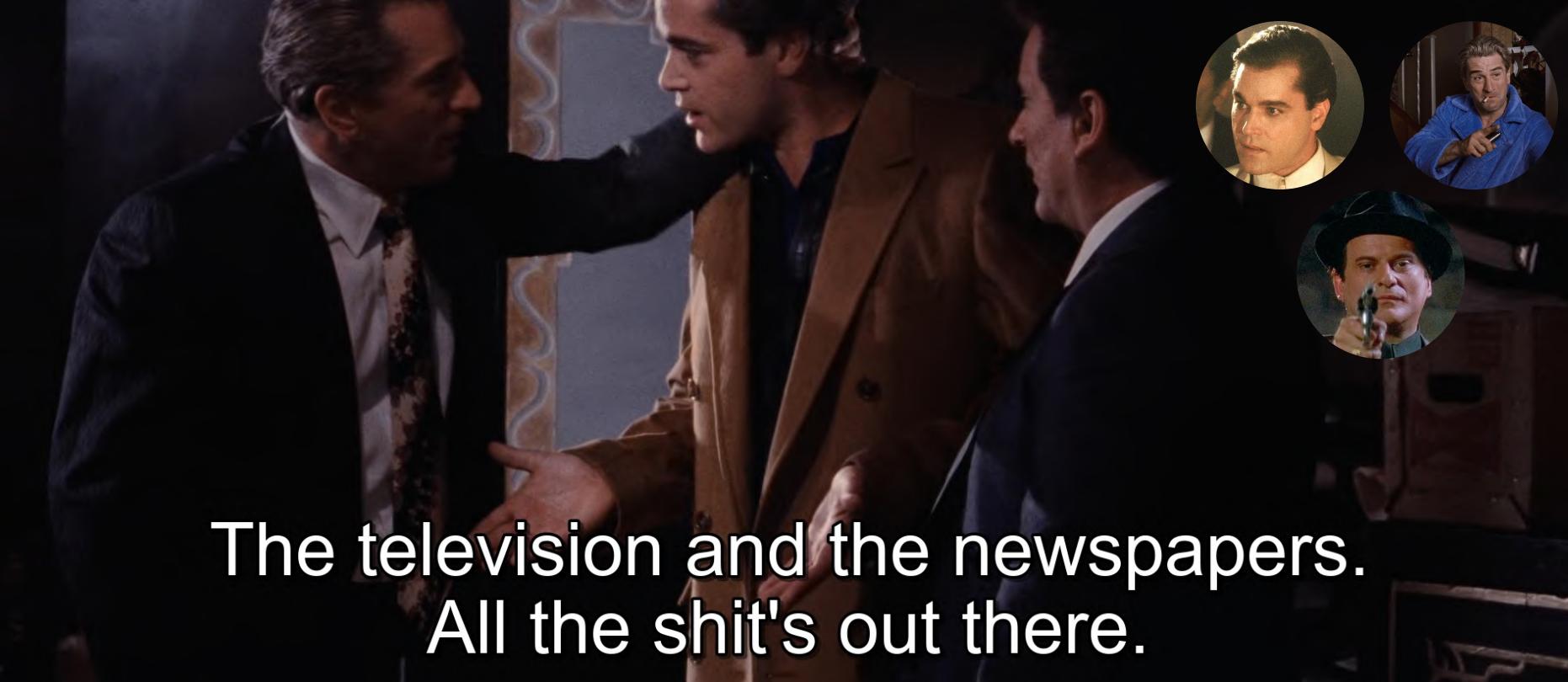
Try Pitch

Peak of Sadness

GoodFellas



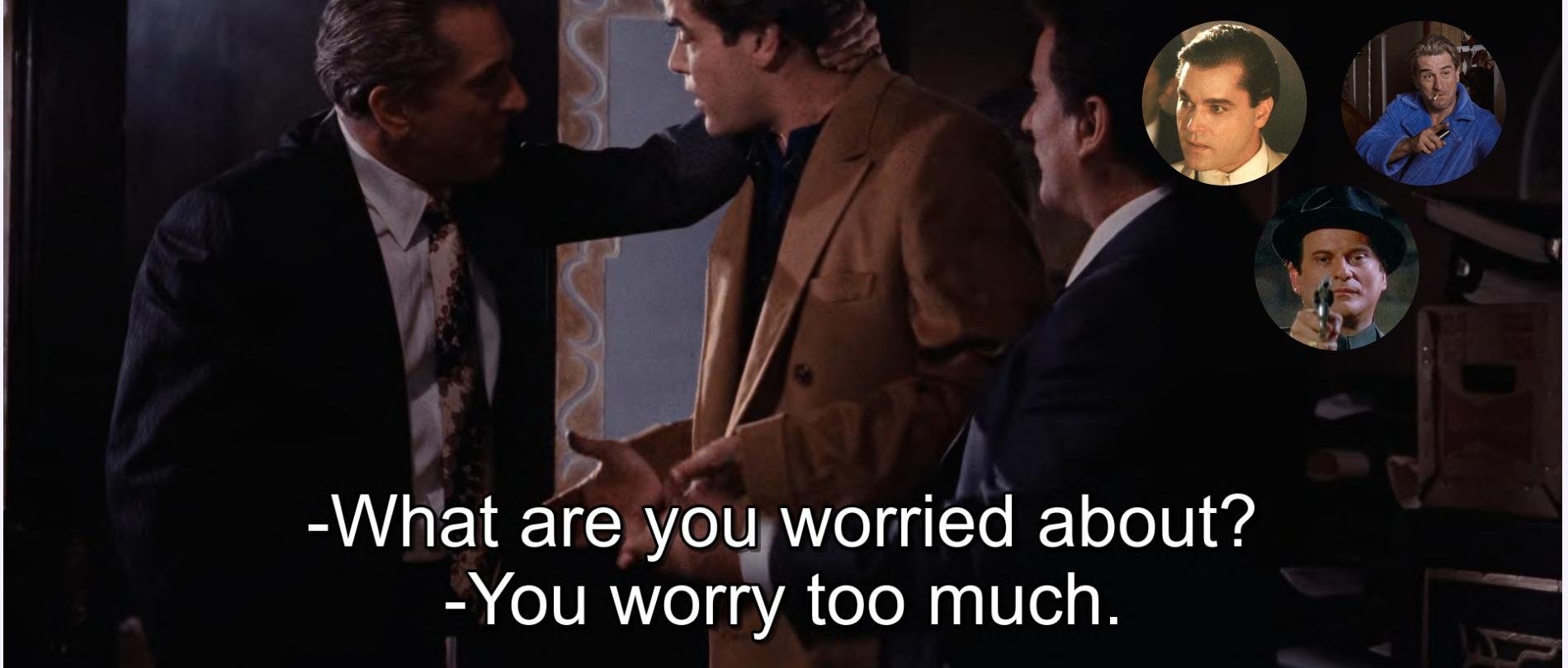
It was just a matter of time
before they got to Stacks.



The television and the newspapers.
All the shit's out there.



-What happened with Stacks?
-Don't worry about that.



-What are you worried about?
-You worry too much.

Similar* scenes between movies

CASINO

GoodFellas

Casino chunks	Casino-Time Range	Goodfellas chunks	Goodfellas-Time Range	Similarity
26	01:26:29 - 01:29:49	10	00:25:57 - 00:28:44	0.999639
45	02:31:27 - 02:34:52	25	01:07:42 - 01:10:20	0.999496
16	00:52:10 - 00:55:37	17	00:45:26 - 00:48:10	0.999472
50	02:48:38 - 02:51:57	7	00:17:40 - 00:20:25	0.999290
15	00:48:50 - 00:52:10	36	01:38:17 - 01:40:51	0.999278
25	01:23:01 - 01:26:20	22	00:59:16 - 01:02:01	0.999128
44	02:28:02 - 02:31:27	15	00:39:48 - 00:42:36	0.998966
42	02:21:11 - 02:24:32	31	01:24:13 - 01:27:00	0.998832
30	01:40:06 - 01:43:31	22	00:59:16 - 01:02:01	0.998806
15	00:48:50 - 00:52:10	30	01:21:28 - 01:24:13	0.998694

Similar* scenes between movies

CASINO

GoodFellas

Chunk	Time Range	fear	joy	anger	sadness	Chunk	Time Range	fear	joy	anger	sadness
26	01:26:29 - 01:29:49	34.782609	14.492754	30.434783	20.289855	10	00:25:57 - 00:28:44	34.042553	14.893617	29.787234	21.276596
45	02:31:27 - 02:34:52	26.086957	10.869565	50.000000	13.043478	25	01:07:42 - 01:10:20	25.490196	11.764706	50.980392	11.764706
16	00:52:10 - 00:55:37	41.666667	15.277778	22.222222	20.833333	17	00:45:26 - 00:48:10	42.000000	14.000000	22.000000	22.000000
50	02:48:38 - 02:51:57	41.666667	13.888889	27.777778	16.666667	7	00:17:40 - 00:20:25	41.304348	15.217391	28.260870	15.217391
15	00:48:50 - 00:52:10	21.875000	15.625000	32.812500	29.687500	36	01:38:17 - 01:40:51	23.529412	14.705882	32.352941	29.411765
25	01:23:01 - 01:26:20	33.333333	5.263158	42.105263	19.298246	22	00:59:16 - 01:02:01	33.333333	6.666667	40.000000	20.000000
44	02:28:02 - 02:31:27	32.352941	16.176471	39.705882	11.764706	15	00:39:48 - 00:42:36	33.333333	16.666667	37.500000	12.500000
42	02:21:11 - 02:24:32	39.393939	9.090909	27.272727	24.242424	31	01:24:13 - 01:27:00	40.000000	10.000000	25.000000	25.000000
30	01:40:06 - 01:43:31	34.000000	6.000000	42.000000	18.000000	22	00:59:16 - 01:02:01	33.333333	6.666667	40.000000	20.000000
15	00:48:50 - 00:52:10	21.875000	15.625000	32.812500	29.687500	30	01:21:28 - 01:24:13	23.809524	14.285714	33.333333	28.571429

Similar* characters between movies

CASINO	GoodFellas		
Casino character	Goodfellas Character		Similarity
NICKY		 TOMMY	0.995234
GAGGI		JIMMY	0.994413
GINGER		 TOMMY	0.991113
ACE		 HENRY	0.987514
ACE		JIMMY	0.986548
LESTER		HENRY	0.982325
MARINO		JIMMY	0.979159
NICKY		 JIMMY	0.960231
LESTER		 KAREN	0.959776
ACE		 TOMMY	0.956709

Ten **most similar** characters

CASINO	GoodFellas		
Casino character	Goodfellas Character		Similarity
JENNIFER		TOMMY	0.654846
MARINO		KAREN	0.744199
PISCANO		KAREN	0.762798
NICKY		 KAREN	0.822653
JENNIFER		JIMMY	0.830106
GAGGI		KAREN	0.839462
GINGER		 KAREN	0.866172
LESTER		 TOMMY	0.866394
PISCANO		HENRY	0.882601
JENNIFER		HENRY	0.899176

Ten **least similar** characters



THE END