Predicting movie-elicited emotions from dialogue in screenplay text: A study on "Forrest Gump"

Benedetta Iavarone[⋆ ⋄], **Felice Dell'Orletta**[⋄]
[⋆] Scuola Normale Superiore, Pisa
[⋄]ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
benedetta.iavarone@sns.it, felice.dellorletta@ilc.cnr.it

Abstract

We present a new dataset of sentences[1] extracted from the movie Forrest Gump, annotated with the emotions perceived by a group of subjects while watching the movie. We run experiments to predict these emotions using two classifiers, one based on a Support Vector Machine with linguistic and lexical features, the other based on BERT. The experiments showed that contextual embeddings are effective in predicting human-perceived emotions.

1 Introduction

Emotional intelligence, described as the set of skills that contributes to the accurate appraisal, expression and regulation of emotions in oneself and in others (Salovey and Mayer, 1990), is recognised to be one of the facets that make us humans and the fundamental ability of human-like intelligence (Goleman, 2006). Emotional intelligence has played a crucial role in numerous applications during the last years (Krakovsky, 2018), and being able to pinpoint expressions of human emotions is essential to advance further in technological innovation. Emotions can be identified in many sources, among which there are semantics and sentiment in texts (Calefato et al., 2017). In NLP, Sentiment Analysis already boasts many state-of-the-art tools that can accurately predict or classify the polarity of a text. However, real applications often need to go beyond the dichotomy positive-negative and identify the emotional content of a text with a finer granularity. Nevertheless, the task of predicting a precise emotion from text brings many challenges, mostly because there is a need of context: emotions can't be easily understood in isolation, as they are conveyed by a complex of explicit (e.g. speech) and implicit (e.g. gesture and posture) behavioural cues. Still, there has been an increasing interest in research for text-based emotion detection (Acheampong et al., 2020). In this work, we study how textual information extracted from the screenplay of a movie can be used to predict the emotions perceived by a group of people during the view of the movie itself. We create a new dataset of sentences extracted from the screenplay, annotated with six different perceived emotions and their perceived intensity and create a binary classification task to predict emotional elicitation during the view of the movie. We use two predicting models, with different kind of features that capture diverse language information. We determine which model and which kind of features are the best for predicting the emotions perceived by the subjects.

2 Data

Our dataset was retrieved from *studyforrest*[2], a research project centered around the use of the movie *Forrest Gump*. The project repository contains data contributions from various research groups, divided in three areas: (*i*) behavior and brain function, (*ii*) brain structure and connectivity, and (*iii*) movie stimulus annotations. We focused on the latter, retrieving two types of data: the speech present in the movie and the emotions that the vision of the movie elicited in a group of subjects. As for the speech, each screenplay line pronounced by the characters is transcribed in sentences and associated with two timestamps in terms of tenths of a second t_{begin} and t_{end}, that respectively indicate the moment of the movie in which the character starts talking and the moment in which they stop. Emotional data comes from the contribution to the project given by Lettieri et al. (2019). A group of 12 subjects was asked to watch the movie and

[1]Data can be downloaded at www.italianlp.it/dataset_release.zip

[2]http://studyforrest.org/

Subject	Happiness	Surprise	Fear	Sadness	Anger	Disgust	Neutral	Emotion
1	592	172	101	557	111	166	22	876
2	628	87	83	539	120	42	61	837
3	345	471	212	340	123	37	30	868
4	274	179	137	255	119	133	276	622
5	244	84	98	224	83	6	305	593
6	496	92	147	264	60	13	113	785
7	277	255	88	132	88	23	286	612
8	357	218	119	305	103	77	231	667
9	299	389	15	147	109	22	312	586
10	213	125	81	255	60	0	377	521
11	352	320	116	307	150	30	120	778
12	180	36	22	149	34	25	526	372
Total	4257	2428	1219	3474	1160	574	2659	8117

Table 1: Emotions distribution in the dataset.

report the emotions they were experiencing during the vision, among a list of six emotions (happiness, surprise, fear, sadness, anger, disgust). Emotion reporting was performed by pressing the keys of a keyboard, with which subjects could indicate the emotion they were experiencing and its intensity, within a range from 0 (no emotion) to 100.

2.1 Data creation

Emotional data was collected from a continuous output $\mathbf{z} = (z_1, z_2, ..., z_n)$ from the keyboard, such that each z_i corresponds to an increment of 0.1 seconds in the playing time of the movie ($z_i = 0.1$, $z_{i+1} = 0.2$, $z_{i+2} = 0.3$, ...). Each z_i is associated to a list $x_{i1}, x_{i2}, ..., x_{ij}$, with $x_j \in [0, 100]$ and $j \in$ [*happiness, surprise, fear, sadness, anger, disgust*], where each x_j indicates the intensity that one emotion assumes at a given timestamp. For our purpose, this information was too detailed and it could not be mapped to textual data properly, thus we proceeded to resample emotional information. We generated new timestamps $\mathbf{s} = (s_1, s_2, ..., s_m)$, such that each s_i corresponds to the sum of 20 consecutive z_i, thus to an increment of 2 seconds in the playing time of the movie. Each s_i is associated to a new list of emotional values, where each new value is the average of the values associated to the summed z_i.

After resampling, we aligned the text to emotional data. As one of our aims is to determine how much text is needed for accurate emotion prediction, we considered three progressively larger time windows for each s_k, such that $window_i = [s_k - m, s_k]$, where $m = (2, 4, 6)$. For each sentence, we retrieve its t_{end} and align the sentence verifying if $s_k - m \leq t_{end} \leq s_k$, thus checking if the moment in which the sentence ends falls within the given time window. In this way, the larger the time window, the larger the amount of text that gets aligned with a specific timestamp. With this process, we created three different datasets, one for each time window. We then removed all the lines in which no text was aligned to s_k. For each dataset, we end up having 898 timestamps associated with a line of text and 6 emotion declarations for each of the 12 subjects.

2.2 Data statistics and data selection

We first looked at the distribution of our data, examining how many times each subject declared a specific emotion. Whenever the subject assigned a value different than zero to a certain emotion, we considered that emotion as present at a given timestamp, regardless of its intensity. If all 6 emotions were zero at the same time (all $x_j = 0$), we assigned to that case the class *neutral*. Furthermore, if any given emotion was declared (at least one $x_j \neq 0$), we assigned to that case the class *emotion*, to indicate a generic emotional response.

As shown in Table 1, the most represented emotions in the dataset are happiness and sadness, while the others are underrepresented. Table 1 also shows that emotions distribution is quite uneven among the different subjects, as there were some subjects that declared emotions frequently and others that entered fewer declarations. This is due to the fact that emotive phenomena are strongly subjective, meaning that emotion processing is specific to each person and that everyone experiences emotions at a different granularity (Barrett, 2006). To account for this factor, we measured the level of agreement between the 12 subjects using *Fleiss' Kappa*. Table 2 reports the percentage of agreement for each emotion in the data. The lowest agreement was found on surprise and disgust. As disgust is also the less declared emotion, it is fair to assume

Emotion	Agreement
happiness	0.32
surprise	0.14
fear	0.41
sadness	0.31
anger	0.42
disgust	0.17

Table 2: Annotators agreement (Fleiss' Kappa) on all emotions

that the movie does not contain many moments that elicit this emotion in the subjects. On the other side, the strongest agreement is found on fear and anger, showing that these emotions are evoked in specific scenes of the movie and that subjects had a similar emotional response to those scenes. In Table 3 we report examples of sentences on which the subjects agreed the most, for all six emotions. For every emotion, there are many sentences on which a large number of subjects agreed, meaning that there were various moments of the movie that elicited the same emotions in the subjects. In the case of disgust, the highest level of agreement was achieved at 8 subjects, only on one sentence. There were no other sentences for which 8 subjects (or more) agreed. This is justified by the fact that disgust is the less represented emotion in the data.

Given the information on the agreement and on emotions distribution, we decided not to examine underrepresented emotions directly, even if their agreement was strong (i.e. surprise). In order to still account for underrepresented emotions, we relied on the general class *emotion*. Hence we assessed three different scenarios: (*i*) the presence of any kind of emotion (at least one $x_j \neq 0$), (*ii*) the presence of happiness ($x_{happiness} \neq 0$) and (*iii*) the presence of sadness ($x_{sadness} \neq 0$). Furthermore, we decided to conduct our experiments only on two subjects, subject 4 and subject 8. We focused on these specific subjects as they declared all emotions evenly, without neglecting any of them, and because the number of declarations for each emotion was quite similar between the two subjects.

3 Emotions prediction

We evaluated the three scenarios described in 2.2 in contrast to the absence of any emotion (all $x_j = 0$), producing three binary classification tasks. We relied on two sets of features: automatically extracted linguistic and lexical features, and contextual word embeddings from a language model.

Emotion	N subjs	Text
happiness	12	I had never seen anything so beautiful in my life. She was like an angel.
surprise	11	Jenny! Forrest!
fear	12	(into radio) Ah, Jesus! My unit is down hard and hurting! 6 pulling back to the blue line, Leg Lima 6 out! Pull back! Pull back!
sadness	12	Bubba was my best good friend. And even I know that ain't something you can find just around the corner. Bubba was gonna be a shrimpin' Boat captain, But instead he died right there by that river in Vietnam.
anger	12	Are you retarded, Or just plain stupid? Look, I'm Forrest Gump.
disgust	8	You don't say much, do you?

Table 3: Examples of sentences on which subjects agreed the most, for all emotions.

3.1 Prediction with linguistic and lexical features

For the first set of features, sentences were first POS tagged and parsed using UDPipe (Straka and Straková, 2017). We extracted a wide set of features, like the ones described in Brunato et al. (2020). These features capture various linguistic phenomena, that range from raw information to information related to the morpho-syntactic and syntactic structure of the sentence (rows 1, 2 and 3 in Table 4, hereafter *linguistic* features). Additionally, we extracted other features that are able to capture some lexical information (row 4 in Table 4, hereafter *lexical* features), as they identify set of characters or words that appear more frequently within a sentence. We trained two SVM models, one on the linguistic features (*SVMling*), one on the lexical features (*SVMlex*). We trained the models with a linear kernel and standard parameters, performing 10-cross-fold validation to evaluate the models accuracy.

3.2 Prediction with language model

For the second set of features, we relied on BERT (Devlin et al., 2019), a Neural Language Model that encodes contextual information. We retrieved the pre-trained base model and fine tuned it on our data. The pre-trained BERT model already includes a lot of information about the language, as it has already been trained on a large amount of data. By fine tuning it on our data, we are able to exploit the information already acquired by the model and use it for our task. We performed differ-

ent fine tuning stages, then used the so fine-tuned models to perform the binary classification task on our data. We evaluated model accuracy using 10 cross-fold-validation. Specifically, we tested three different fine tuning approaches: (1) original data (*BERTorig*), (2) oversampled data to balance the neutral class (*BERTover*), (3) oversampled data + transfer learning tuning (*BERTtransf*). In the case of (3), we first fine tuned the model on data different than ours but conceived for a similar task. Notably, we relied on data created for SemEval-2018 Task 1E-c (Mohammad et al., 2018), containing tweets annotated with 11 emotion classes. After this first tuning, we tuned the model again on our oversampled data and proceeded with the classification task.

4 Results and discussion

Figure 1 shows the accuracy scores for all the models, for both subjects and the three datasets. In all cases, the baseline was determined with a majority classifier. The results appear similar for both subjects.

SVM models are always outperformed by BERT ones. In any case, *SVMling* is the model that gave the lowest performance, remaining below or around the baseline value. On the contrary, *SVMlex* tends to bring a higher performance, despite remaining close to the baseline in most cases. On one side, this is due to the fact that features that look at the raw, morpho-syntactic and syntactic aspects of text, do not encode any relevant information regarding the emotional cues in the text. *SVMlex* always performs better than *SVMling* because lexical features look at patterns of words and characters that are repeated in the input text and thus record information about the lexicon of the dataset. However, as our dataset is too small, it is hard for the model to retrieve the same lexical patterns in both the training and test set.

BERT models outperform the SVM ones in both happiness and sadness prediction. In the case of emotion prediction, BERT models obtain very good results only on the 6 seconds dataset. This is due to the fact that, in this case, we have flattened all emotions into a single category, thus it may be difficult for the model to distinguish between general emotionally charged sentences and those that are not perceived as emotionally charged. When emotions are specific and clearly separated, as in happiness and sadness cases, BERT is able to infer the perceived emotions even from small amounts of text (2 and 4 seconds datasets). *BERTover* and *BERTtransf* tend to give better performances than what happens with *BERTorig*. In the case of *BERTover*, there is a very slight difference in the prediction of happiness and sadness, as in these cases the classes to be predicted were distributed quite evenly. In the case of emotion prediction, the model is helped by the higher representation of the neutral class. With *BERTtransf*, the performances stay in line with the ones obtained with the bare oversampling. Fine tuning the model on similar data did not add any more useful information. This is due to the fact that SemEval data were too distant from the ones of our dataset. Therefore, even though the task is similar to ours, the input text is too different from our sentences to actually make a huge difference for the prediction. We also tried another form of transfer learning, tuning the model on one subject and testing it on the other one. However, the results were too low and we did not report them. This is because emotion perception is a very personal phenomenon and it cannot be easily generalised to different individuals.

To further evaluate the results, we computed the percentage of agreement between the two models that overall had the best performances, *BERTover*

Level of Annotation	**Feature**
Raw Text	Sentence length
	Word length
	Type/Token Ratio for words and lemmas
POS tagging	Distibution of POS
	Lexical density
	Inflectional morphology of lexical verbs and auxiliaries (Mood, Number, Person, Tense and VerbForm)
Dependency Parsing	Depth of the whole syntactic tree
	Average length of dependency links and of the longest link
	Average length of prepositional chains and distribution by depth
	Clause length (n. tokens/verbal heads)
	Order of subject and object
	Distribution of verbs by arity
	Distribution of verbal heads and verbal roots
	Distribution of dependency relations
	Distribution of subordinate and principal clauses
	Average length of subordination chains and distribution by depth
	Relative order of subordinate clauses
Lexical Patterns	Bigrams, trigrams and quadrigrams of characters, words and lemmas

Table 4: Linguistic and Lexical Features.

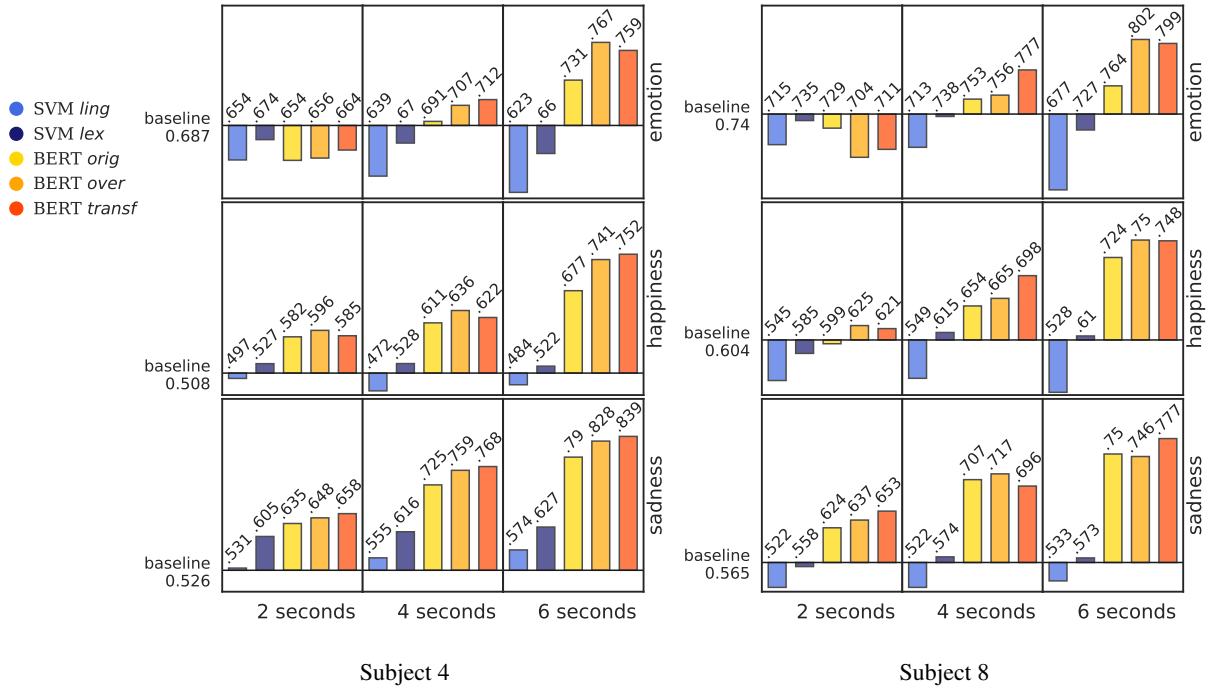

Figure 1: Performances (accuracy) of SVM and BERT models in the prediction of emotion, happiness of sadness, for every timespan window, and for both subject 4 and subject 8.

	subject 4			subject 8		
	2sec	4sec	6sec	2sec	4sec	6sec
emo.	82.75	83.63	85.82	82.03	87.8	90.44
hap.	70.77	72.64	79.78	76.26	72.31	79.67
sad.	82.53	85.93	87.47	80.44	79.45	85.05

Table 5: Agreement (%) between *BERTover* and *BERTtransf* predictions.

and *BERTtransf*. We defined agreement as the percentage of sentences for which the models gave the same output during the classification task. Table 5 reports the results for emotion, happiness and sadness, for every timespan window, and for both subjects 4 and subject 8. The agreement is quite high in all cases, and it tends to get stronger with the amount of text on which models are trained (i.e. 6 seconds). A higher level of agreement indicates that the models have similar behaviour, thus making the same mistakes in the classification task. The lowest levels of agreement are encountered on the classification of happiness, showing that the two models work differently in this part of the task. Indeed, both *BERTover* and *BERTtransf* obtain high performances in predicting happiness, but the fact that their agreement is lower suggests that they differ in the mistakes they make in the classification. We may exploit this information to create systems that combine different classifiers, actually enhancing the classification accuracy. By doing this, it is possible to compare the cases in which two or more classifiers agree and the cases in which they make mistakes, thus choosing the best classification output accordingly.

5 Conclusion

In this paper, we presented a dataset of sentences extracted from the movie Forrest Gump, annotated with the emotions that a group of subjects perceived while watching the movie, and we studied how to predict these emotions. To do so, we retrieved different kinds of features from the sentences pronounced by the characters of the movie. We showed that contextual embeddings extracted from the sentences can accurately predict specific emotions, even if the amount of text used for the prediction is very little. Instead, when predicting generic emotional elicitation, a larger amount of text is required for an accurate prediction. We also show that lexical, morpho-syntactic and syntactic aspects of the sentences cannot be used to infer emotional elicitation during the view of the movie.

As emotional response is directly correlated with brain activity, we plan to add fMRI images recorded during the vision of the movie to the contextual embedding we extracted. In this way, we could verify if brain images can help to increase the accuracy in the prediction of perceived emotions.

Acknowledgments

We thank MoMiLab research group of IMT Lucca for having shared with us the data they collected on human-perceived emotions. Furthermore, we are grateful to the *studyforrest* project and all its contributors.

References

[Acheampong et al.2020] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, page e12189.

[Barrett2006] Lisa Feldman Barrett. 2006. Valence is a basic building block of emotional life. *Journal of Research in Personality*, 40(1):35–55.

[Brunato et al.2020] Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.

[Calefato et al.2017] Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. Emotxt: a toolkit for emotion recognition from text. In *2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[Goleman2006] Daniel Goleman. 2006. *Emotional intelligence*. Bantam.

[Krakovsky2018] Marina Krakovsky. 2018. Artificial (emotional) intelligence.

[Lettieri et al.2019] Giada Lettieri, Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Paolo Papale, Monica Betta, Pietro Pietrini, and Luca Cecchetti. 2019. Emotionotopy in the human right temporo-parietal cortex. *Nature communications*, 10(1):1–13.

[Mohammad et al.2018] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

[Salovey and Mayer1990] Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.

[Straka and Straková2017] Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.