

1 Supplementary Material

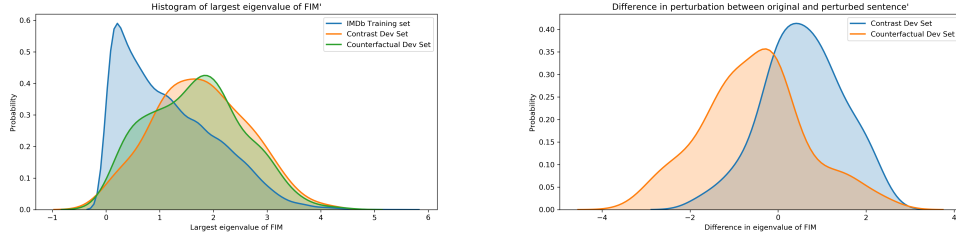


Figure 1: a) Histogram of eigenvalues of dev sets of original IMDb examples, contrast sets and counterfactual examples. The significant overlap between the three distributions indicates that the counterfactual and contrast set examples might not be as difficult as previously believed. The tail end of all three distributions contain the difficult examples. b) Distribution of difference in largest eigenvalue of FIM of the original and the perturbed sentence in contrast set and counterfactual examples. With a mean near 0, these perturbations are not necessarily more difficult for the model.

Table 1: Perturbing an easy example on the IMDb dataset: The first row represents original sentence. As we can see here, most perturbations are ineffective in changing the FIM eigenvalue and thus the difficulty of the example. Despite multiple substitutions (last row), we are only able to achieve a modest increase in FIM score, indicating the resilience to perturbations.

Perturbed sentiment	Word substitutions
Negative → Negative easy example ($\lambda_{max} = 0.0007$)	Probably the worst Dolph film ever. There’s nothing you’d want or expect here. Don’t waste your time. Dolph plays a miserable cop with no interests in life. His brother gets killed and Dolph tries to figure things out. The character is just plain stupid and stumbles around aimlessly. Pointless.
Negative → Negative minor change in FIM value ($\lambda_{max} = 0.0008$)	Probably the worst Dolph film ever. There’s nothing → everything you’d want or expect here. Don’t waste your time. Dolph plays a miserable cop with no interests in life. His brother gets killed and Dolph tries to figure things out. The character is just plain stupid and stumbles around aimlessly. Pointless.
Negative → Negative minor change in FIM value ($\lambda_{max} = 0.0005$)	Probably the worst Dolph film ever. There’s nothing you’d want or expect here. Don’t waste your time. Dolph plays → portrays a miserable cop with no interests in life. His brother gets killed and Dolph tries to figure things out. The character is just plain stupid and stumbles around aimlessly. Pointless.
Negative → Negative minor change in FIM value ($\lambda_{max} = 0.0007$)	Probably the worst Dolph film ever. There’s nothing you’d want or expect here. Don’t waste your time. Dolph plays → portrays a miserable cop with no interests in life. His brother gets killed and Dolph tries → attempts to figure things out. The character is just plain stupid and stumbles around aimlessly. Pointless.
Negative → Negative Significant increase in FIM ($\lambda_{max} = 1.56$)	Probably the best Dolph film ever. There’s everything you’d want or expect here. Spend your time. Dolph portrays a miserable cop with lots of interests in life. His brother gets killed and Dolph attempts to figure things out. The character is just plain amazing .

Table 2: Perturbing a difficult example on the IMDb dataset: The first row represents original sentence. Unlike easy examples, difficult examples tend to have a mixture of positive and negative traits. Furthermore, a minor perturbation like removal of a sentence (row 1) or substitution of a word (row 2) causes a significant drop in FIM eigenvalue. The small FIM score indicates that the perturbed review is not difficult for the classifier.

Perturbed sentiment	Word substitutions
Negative → Negative difficult example ($\lambda_{max}=5.47$)	It really impresses me that it got made. The director/writer/actor must be really charismatic in reality. I can think of no other way itd pass script stage. What I want you to consider is this...while watching the films I was feeling sorry for the actors. It felt like being in a stand up comedy club where the guy is dying on his feet and your sitting there, not enjoying it, just feeling really bad for him coz hes of trying. Id really like to know what the budget is, guess it must have been low as the film quality is really poor. I want to write 'the jokes didn't appeal to me'. but the reality is for them to appeal to you, you'd have to be the man who wrote them. or a retard. So imagine that in script form...and this guy got THAT green lit. Thats impressive isn't it?
Negative → Negative Significant change in FIM ($\lambda_{max}=0.640$)	It really impresses me that it got made. The director/writer/actor must be really charismatic in reality. I can think of no other way itd pass script stage. What I want you to consider is this...while watching the films I was feeling sorry for the actors. It felt like being in a stand up comedy club where the guy is dying on his feet and your sitting there, not enjoying it, just feeling really bad for him coz hes of trying. Id really like to know what the budget is, guess it must have been low as the film quality is really poor. I want to write 'the jokes didn't appeal to me'. but the reality is for them to appeal to you, you'd have to be the man who wrote them. or a retard. So imagine that in script form...and this guy got THAT green lit. Thats impressive isn't it?
Negative → Negative Significant change in FIM ($\lambda_{max}=0.445$)	It really impresses me that it got made. The director/writer/actor must be really charismatic in reality. I can think of no other way itd pass script stage. What I want you to consider is this...while watching the films I was feeling sorry for the actors. It felt like being in a stand up comedy club where the guy is dying on his feet and your sitting there, not enjoying it, just feeling really bad for him coz hes of trying. Id really like to know what the budget is, guess it must have been low as the film quality is really poor. I want to write 'the jokes didn't appeal to me'. but the reality is for them to appeal to you, you'd have to be the man who wrote them. or a retard. So imagine that in script form...and this guy got THAT green lit. Thats impressive → weird isn't it?

Table 3: Counterfactual perturbations cause reduction in difficulty: The original example is more difficult than the perturbed counterfactual example because of strong giveaway words like "amazing". The subtle changes to make the sentence positive like changing the amount or the rating almost makes no difference. Instead, the use of a strong word "amazing" makes the sentence extremely easy for the model to classify as positive. The heavy reliance on a single word for the positive example makes this much easier to classify than the original sentence which used the word "odd" (a word that is not necessarily negative) as a negative sentiment.

Perturbed sentiment	Word substitutions
Negative → Negative difficult example ($\lambda_{max}=3.42$)	Definitely an odd debut for Michael Madsen. Madsen plays Cecil Moe, an alcoholic family man whose life is crumbling all around him. Cecil grabs a phone book, looks up the name of a preacher, and calls him in the middle of the night. He goes to the preacher's home and discusses his problems. The preacher teaches Cecil to respect the word of God and have Jesus in his heart. That makes everything all better. Ahh...if only everything in life were that easy. The fact that this "film" looks as if it was made with about \$500 certainly doesn't help. 1/10
Positive → Positive Significant change in FIM ($\lambda_{max}=0.640$)	Definitely an amazing debut for Michael Madsen. Madsen plays Cecil Moe, an alcoholic family man whose life is crumbling all around him. Cecil grabs a phone book, looks up the name of a preacher, and calls him in the middle of the night. He goes to the preacher's home and discusses his problems. The preacher teaches Cecil to respect the word of God and have Jesus in his heart. That makes everything all better. Ahh...if only everything in life were that easy. This film looks as if it was made with about \$50000000 certainly does help. 10/10

Table 4: Difficult examples on the IMDB counterfactual dataset. Here the first row is the original sentence and the next row is the counterfactual sentence. The counterfactual sentence is easier than the original sentence. Note the original example relied more on words like “unlucky”, “doesn’t”, and “absolutely” during classification. “unlucky” and “doesn’t” are associated with more negative sentences and thus the counterfactual example is much easier for the model along with very negative words like “terrible” and “boring”.

Perturbed sentiment	Word substitutions
Positive → Positive diffi- cult example $(\lambda_{max} = 4.04)$	An excellent movie about two cops loving the same woman. One of the cop (P��rier) killed her, but all the evidences seems to incriminate the other (Montand). The unlucky Montand doesnt know who is the other lover that could have killed her, and P��rier doesnt know either that Montand had an affair with the girl. Montand must absolutely find the killer...and what a great ending! Highly recommended.
Negative → Negative Significant change in FIM $(\lambda_{max} = 0.35)$	A terrible movie about two cops loving the same woman. One of the cop (P��rier) killed her, but all the evidences seems to incriminate the other (Montand). The unlucky Montand doesnt know who is the other lover that could have killed her, and P��rier doesnt know either that Montand had an affair with the girl. Montand must absolutely find the killer...and what a boring ending! I don’t recommend at all.

Table 5: Difficult examples on the IMDB counterfactual dataset. Here the first row is the original sentence and the next row is the counterfactual sentence. Mix of positive and negative words make the sentences difficult for the model.

Perturbed sentiment	Word substitutions
Positive → Positive diffi- cult example $(\lambda_{max} = 4.18)$	Was flipping around the TV and HBO was showing a double whammy of unbelievably horren- dous medical conditions, so I turned to my twin sister and said, "Hey this looks like fun," - truly I love documentaries - so we started watching it. At first I thought Jonni Kennedy was a young man, but then it was explained that due to his condition, he never went through puberty, thus the high voice and smaller body. He was on a crusade to raise money for his cause. He had the most wonderful sense of humor combined with a beautiful sense of spirituality... I cried, watched some more, laughed, got up to get another Kleenex, then cried some more. Once Jonni Kennedy’s "time was up" he flew to heaven to be with the angels. He was more than ready; he had learned his lessons from this life and he was free. I highly recommend this. If you do not fall in love with this guy, you have no heart.
Negative → Negative Significant change in FIM $(\lambda_{max} = 2.45)$	Was flipping around the TV and HBO was showing a double whammy of unbelievably horren- dous medical conditions, so I turned to my twin sister and said, "Hey this looks like fun," - truly I love documentaries - so we started watching it. At first I thought Jonni Kennedy was a young man, but then it was explained that due to his condition, he never went through puberty, thus the high voice and smaller body. He was on a crusade to raise money for his cause. He had the worst sense of humor combined with an ugly sense of spirituality... I nodded off, watched some more, snoozed, got up to get a coffee, then snoozed some more. Once Jonni Kennedy’s "time was up" he flew to heaven to be with the angels. He was more than ready; he had learned his lessons from this life and he was free. I highly recommend you don’t watch this. If you do not fall asleep within the first ten minutes, you have no taste.

Table 6: Difficult examples on the IMDB contrast dataset. Here the first row is the original sentence and the next row is the contrast set sentence. The contrast set sentence is easier than the original sentence. The difficulty in the first sentence is due to words like “irresponsible” and “sloppy”. Thus the negative sentence is much easier for the model.

Perturbed sentiment	Word substitutions
Positive → Positive diffi- cult example $(\lambda_{max}=3.26)$	Here’s another film that doesn’t really need much of a recommendation. It’s a classic comedy, very funny and entertaining and which, of course, ultimately inspired a successful television series which many would say was even better (I enjoy both, personally). For some, it’s hard to warm up to Jack Lemmon and Walter Matthau as Felix Unger and Oscar Madison when they were weaned on the TV show starring Tony Randall and Jack Klugman (or perhaps vice versa). But what we’ve got there in both cases are four good actors who in real life seemed so much like their film counterparts that they managed to make these characterizations their own. It’s Neil Simon’s humorous material that’s key, and where the laughs really originate from. For those who have somehow never heard of THE ODD COUPLE, it’s the story of a neurotic and fussy neat-freak (Lemmon) who is thrown out of a 12-year marriage by his long-suffering wife and takes up residence in the Manhattan apartment of his sloppy and totally irresponsible buddy (Matthau). Pitting these two unlikely roommates together within the same four walls makes for some hugely funny predicaments.
Negative → Negative Significant change in FIM $(\lambda_{max}=0.32)$	Here’s another film that really needs a recommendation to watch. It’s a travesty, unfunny and which, of course, ultimately inspired a unsuccessful television series which many would say was even worse (I hated both, personally). For some, it’s hard to warm up to Jack Lemmon and Walter Matthau as Felix Unger and Oscar Madison when they were weaned on the TV show starring Tony Randall and Jack Klugman (or perhaps vice versa). I am no exception. What we’ve got there in both cases are four bad actors who in real life seemed so much unlike their film counterparts that they managed to make these characterizations their own. It’s Neil Simon’s material that’s the worst, and where the fails really originate from. For those who have somehow never heard of THE ODD COUPLE, it’s the story of a neurotic and fussy neat-freak (Lemmon) who is thrown out of a 12-year marriage by his long-suffering wife and takes up residence in the Manhattan apartment of his sloppy and totally irresponsible buddy (Matthau). Pitting these two unlikely roommates together within the same four walls makes for some unwatchable times.

Table 7: Easy and difficult examples in Contrast Sets

Perturbed sentiment	Word substitutions
Negative → Negative easy example $(\lambda_{max}=0.034)$	This is a pathetic → excellent political satire. No wonder why it was largely ignored in the U.S.: it ridicules our foreign policy and misrepresents what it really is. Another bad → good film from this era, Rendition, was however totally dismissed simply because it showed, accurately, that the U.S. is a war machine bent on torturing, murdering, and maiming civilians in its quest for total world domination. A factually incorrect → correct, bad → good acting, some big stars (John Cusack, Ben Kingsley, Marisa Tomei anyone?) and some scenes of hilarity but they couldn't have made this movie a hit. Thankfully, Americans don't like to hear misrepresentations about anyone, even if they are complicit in mass murder.
Positive → Negative diffi- cult example $(\lambda_{max}=4.132)$	I'm a big Porsche fan, and the car was the best star in this film. Haim, the drug abusing child star of the 80's is amazing → horrible, excellent as per usual, and commenting on back up from minor characters/actors would be pointless; needless to say they were all above average. It's a cool movie as a trip down memory lane into the 80's - with some weird clothes, some good shots of the Colorado backdrop and a very mind stimulating plot. All in all, please watch this unless you hate 80's movies, Corey Haim, or unlike myself, hate old school Porsches (this one in particular looks great) because life's too short to instead watch crappy movies.
Negative → Positive diffi- cult example $(\lambda_{max}=3.98)$	The first film was an okay one, and it is nowhere as good as the wonderful animated classic which I found more poignant and endearing. This sequel is not just inferior, its really bad. Yes the slapstick is too much, the script has its weak spots and the plot is a tad uninspired. Yeah probably the dogs are very cute here, but Eric Idle is dumb as a cow. The film is a pain to look at with any cinematography and eye hurting costumes(especially Cruella's), and the music is sleepy. The acting is mostly very bad, Ioan Gruffudd is appalling → terrible, shocking, excellent, extraordinary and Gerard Depardieu while he has given better performances has little fun as Cruella's accomplice. But the best asset, as it was with the first film, is the amazing Glenn Close in a deliciously over-the-top performance as Cruella, even more evil than she was previously. Overall, poor. 3/10 Bethany Cox

Table 8: Difficult examples for BERT. Sensitive to single word substitutions. In example 1, each word was substituted one at a time.

Perturbed sentiment	Word substitutions
Positive → Negative diffi- cult example $(\lambda_{max}=0.60)$	<p>"The director tries to be Quentin Tarantino, the screenwriters try to be Tennessee Williams, Deborah Kara Unger tries to be Faye Dunaway, the late James Coburn tries to be Orson Welles, Michael Rooker tries to be Gene Hackman, Mary Tyler Moore tries to be Faye Dunaway (older version), Cameron Diaz tries to get out of the frame as quickly as she can (successfully), don't ask about Joanna Going. And they actually pull it off. Eric Stoltz and James Spader try to present their joy with this stuff. It delivers thoughtful → thoughtless, meaningful → meaningless dialog and very little action.

Tulsa is a town with beautiful elevator lobbies, an art deco church by Bruce Goff and a lovely, sprawling mansion by Frank Lloyd Wright. Visit Tulsa, consider watching this movie. It doesn't do the location justice, but still worth it."</p>
Negative → Positive diffi- cult example $(\lambda_{max}=0.58)$	<p>"Many people here say that this show is for kids only. Hm, when I was a kid (approximately 7-9 years old) I watched this show first. It was disgusting → okay for me. I talked with other kids about this and, sure, other shows and know what? This was the measure of disguise, whenever we wanted to emphasize something's silliness (either on TV or anything else) we said 'Uh, just like Power Rangers' and laughed.

And before visiting this site I could not imagine that there actually are fans of MMPR. It was so strange for me that I decided to watch it again and try to understand why people like it. I did not enjoy that viewing. But it dawned upon me: maybe I have not enough imagination? It may be. However this argument is not sufficient for me to rate it more than 1 star."</p>