# Appendix

## A HIGHLIGHTS FOR DIFFERENT METHODS

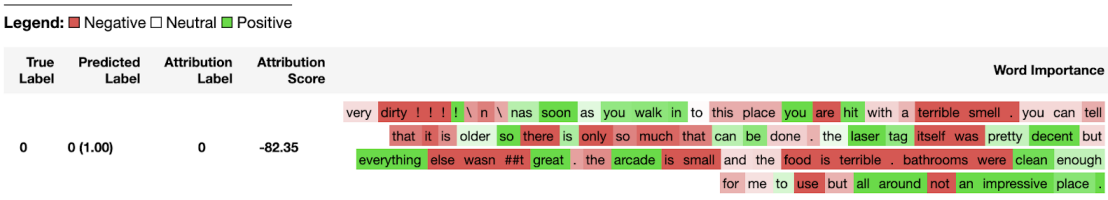


Fig. 1. LIME highlight on Yelp review

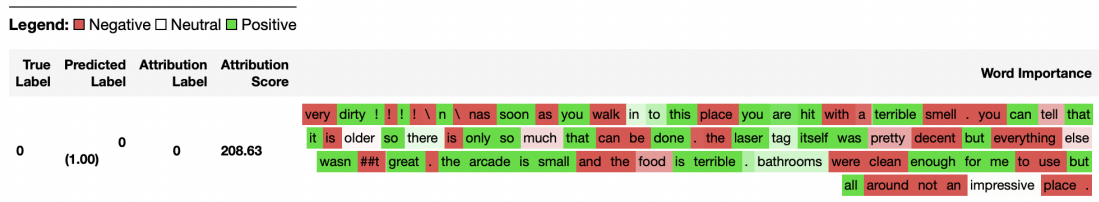


Fig. 2. LIG highlight on Yelp review

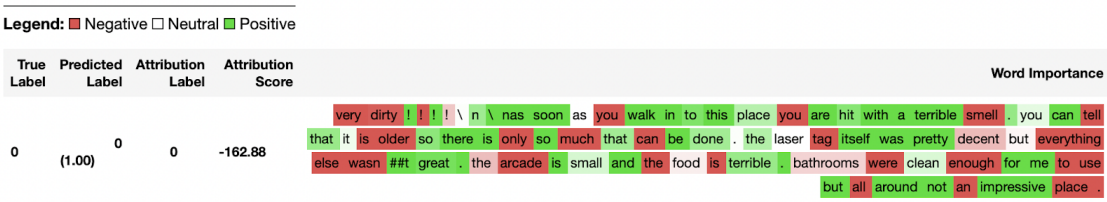


Fig. 3. LxA highlight on Yelp review

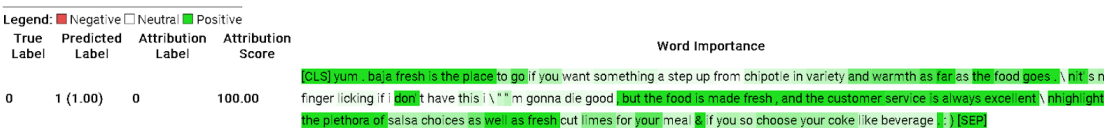


Fig. 4. Attention (positive) highlight on Yelp review

Author's address:

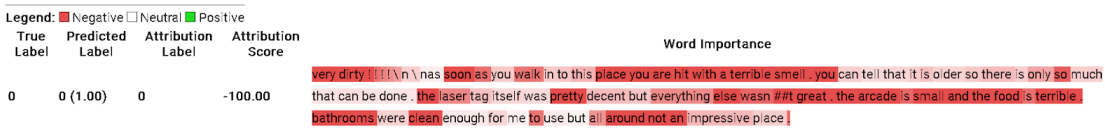


Fig. 5. Attention (negative) highlight on Yelp review

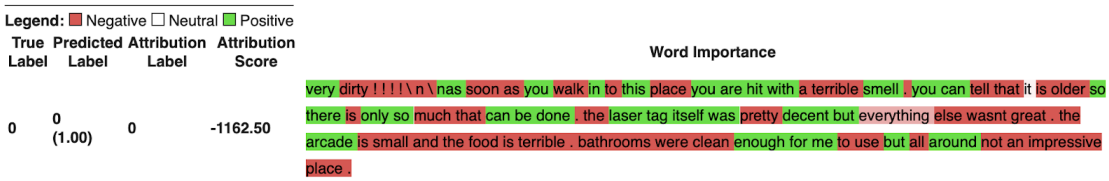


Fig. 6. Logistic Regression highlight on Yelp review

## B LIME HIGHLIGHT-BASED ABLATION

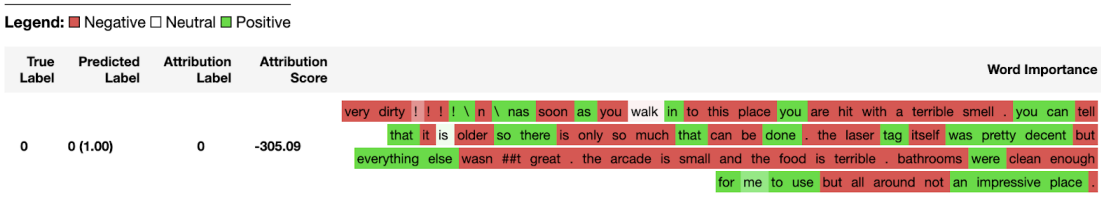


Fig. 7. LIME with 100 point local neighbourhood

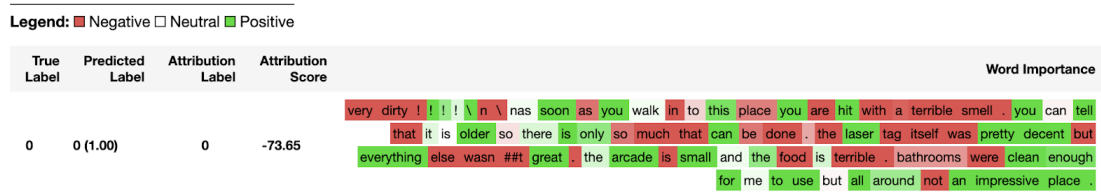


Fig. 8. LIME with 500 point local neighbourhood

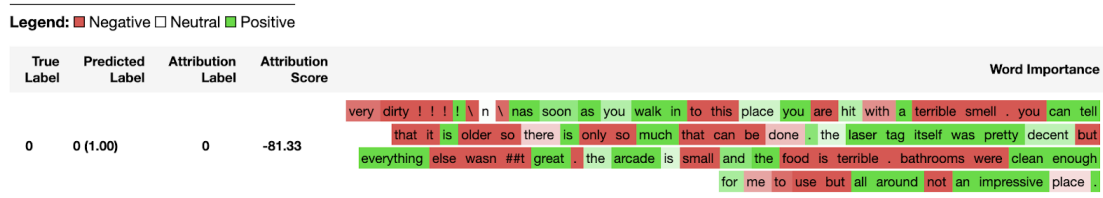


Fig. 9. LIME with 1000 point local neighbourhood

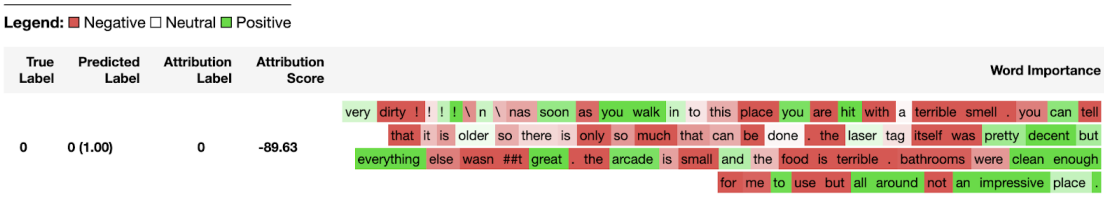


Fig. 10. LIME with 2500 point local neighbourhood

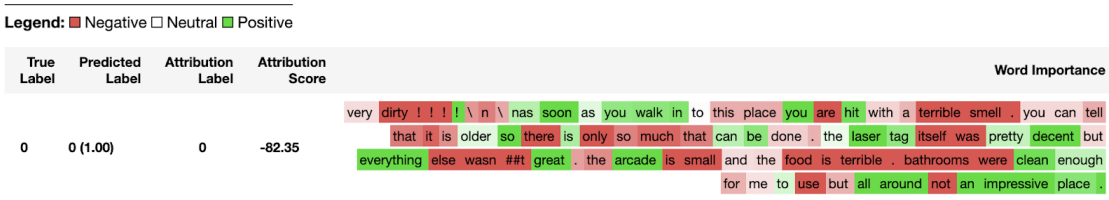


Fig. 11. LIME with 5000 point local neighbourhood

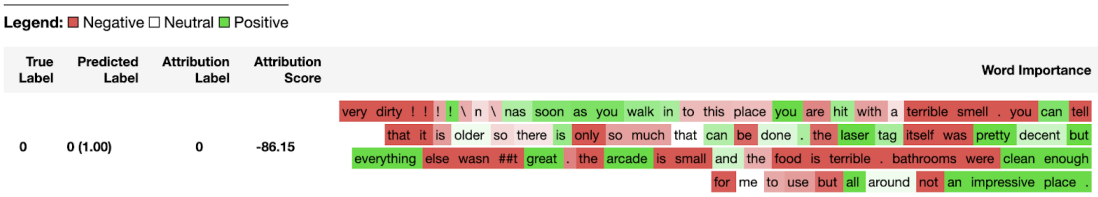


Fig. 12. LIME with 10000 point local neighbourhood

## C ATTENTION BASED VISUALIZATIONS

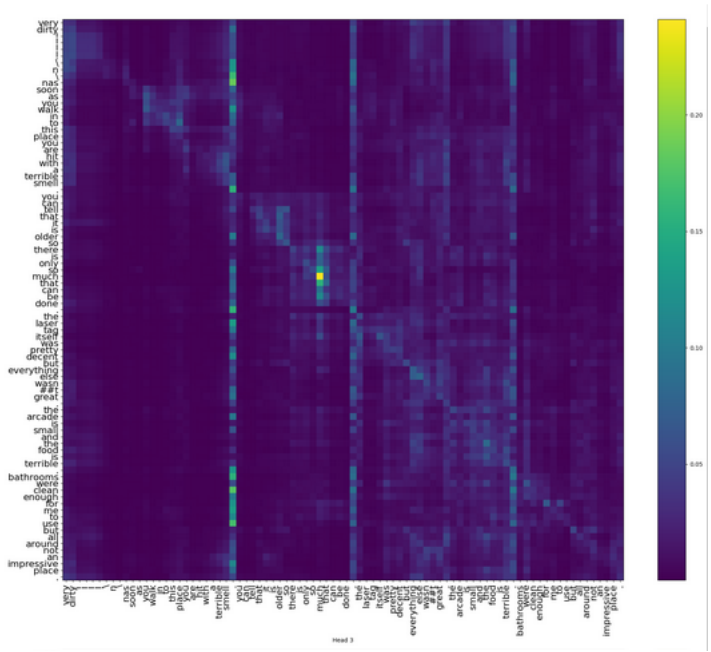

Fig. 13. Heatmap visualization of the Self-Attention weights obtained from the final layer of DistilBERT averaged over the 12 attention heads

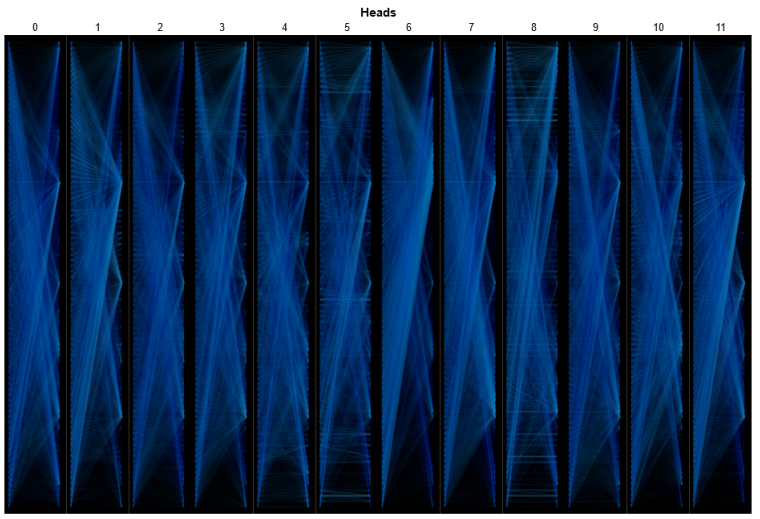


Fig. 14. Bird's-eye view of attention across all attention heads of the final layer for the same example sentence that was used in previous figures (visualization obtained using BertViz).

## D ADDITIONAL EXPERIMENTAL DETAILS

```
DistilBertForSequenceClassification(
  (distilbert): DistilBertModel(
    (embeddings): Embeddings(
      (word_embeddings): Embedding(30522, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (transformer): Transformer(
      (layer): ModuleList(
        (0-5): 6 x TransformerBlock(
          (attention): MultiHeadSelfAttention(
            (dropout): Dropout(p=0.1, inplace=False)
            (q_lin): Linear(in_features=768, out_features=768, bias=True)
            (k_lin): Linear(in_features=768, out_features=768, bias=True)
            (v_lin): Linear(in_features=768, out_features=768, bias=True)
            (out_lin): Linear(in_features=768, out_features=768, bias=True)
          )
          (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (ffn): FFN(
            (dropout): Dropout(p=0.1, inplace=False)
            (lin1): Linear(in_features=768, out_features=3072, bias=True)
            (lin2): Linear(in_features=3072, out_features=768, bias=True)
            (activation): GELUActivation()
          )
          (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        )
      )
    )
  )
  (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
  (classifier): Linear(in_features=768, out_features=2, bias=True)
  (dropout): Dropout(p=0.2, inplace=False)
)
```

Fig. 15. Overview of DistilBERT architecture that we used as our Base model

Some additional details of the computations done for the technique are as follows:

**Base Model**

For the purposes of identifying spurious correlations, we use a DistilBERT Model. DistilBERT is a transformers model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher. Identifying important tokens using various XAI techniques is computationally expensive and time intensive, so we selected a model that minimizes the number of parameters without taking a hit to the accuracy of the model's predictions. The task that we were interested in having our model perform was sentiment analysis. DistilBert models with sentiment analysis classification head are available on Huggingface and have already been fine-tuned for the datasets that we have selected. (For now) The architecture can be observed below.

**Distributed Data Parallel Library**

Techniques like LIME are computationally expensive, so we used three NVIDIA Titan V Volta GPUs. Each contain 12 GB of integrated Graphics memory and operate at a clockspeed of 1200 MHz. To utilize all three GPUs simultaneously, we leveraged the DistributedDataParallel (DDP) library

provided by Pytorch. Using DDP, we spawned a separate process for each GPU. The main process chunked the input dataset into three batches and distributed them between the three processes (on separate GPUs) that were responsible for tokenizing the input and extracting attributions using the specified XAI technique. Once each process was complete, the attributions for each sentence would be amalgamated on the main process where the results could be stored into a pickle file for additional processing. Using DDP on our server, we were able to speed up the process of extracting attributions 3x when compared to using just DataParallel (also provided by Pytorch) or Google Colab (which only provides a single GPU for a limited time).

**Captum Library**

The base model is the same for all of the techniques except for Logistic regression, which itself is a separate model. To collect important scores using: Layer Integrated Gradient (LIG), Layer Gradient x Activation (LxA), and LIME, we utilized the Captum library provided by Pytorch. For all of the Captum libraries, we had to define a (callable) Forward Function to provide as input for the specified attribution class. In all cases, the Forward Function was a softmax of the DistilBERT model that is parameterized by the chunked input data that we described earlier. For all techniques there are 22500 examples from the dataset that are run. LIG runs for 1.3s per sample whereas LxA runs for 1.15s per sample. In contrast, LIME runs at 120 iterations per second for a total of 8 seconds per sample. Whereas LxA and LIG only require evaluating the gradients of the model after a forward pass through the model, LIME requires fitting an additional local & interpretable model to evaluate the effects of perturbing the input. The interpretable model that we used was the SkLearnLasso library (linear model trained with L1 prior as regularizer) from sci-kit learn library. We found experimentally that increasing the size of the number of training examples used to train the local, interpretable model improved the quality of the saliency map generated. Unfortunately, due to resource constraints, we could only run LIME on one example at a time, which means that the local, interpretable model gets retrained for every sentence that we try to obtain a saliency map for. As such, we had to consider the tradeoffs between quality of the saliency map and the run time of computing saliency maps for all of the sentences when determining the hyperparameters for training. Particularly, we choose a 1000 point local neighbourhood for LIME since it captures the tradeoffs well as shown in **Section B.**

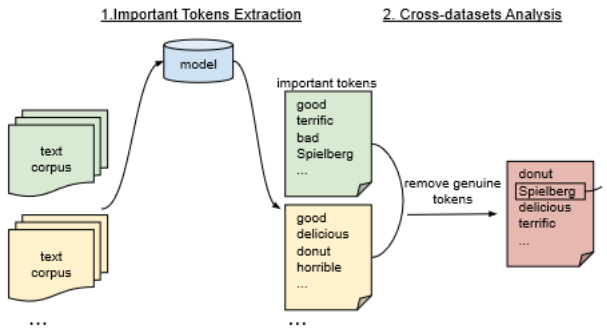


Fig. 16. Overall Framework of our experiments

**Attention Scores**

Wang et al. (Authors that developed the framework that inspired our works) did not specify which of the BERT-based models they used to extract attention scores from. Also, as we have

observed from class, attention scores cannot be directly extracted from the model and immediately be used as a saliency map. To extract the attention scores, we take the transformer block right before the classification head (final layer) as the attention weights in this layer will likely contain the most information regarding the sentiment of each input token. Since the transformer utilizes multi-head attention, we take the average of the attention matrices from each attention head (of which there are 12). This gives us a (seq len x seq len) matrix that contains the self-attention weights for the input sequence. We then take the average of each columnas this score will contain information about each token in the input sequence's relation with every other token in the input sequence. The result is an array that contains a saliency map for the input sequence. Unlike the saliency maps generated by other techniques, we observed that all of the attention scores are strictly non-negative.

**Logistic Regression**

We use Logistic Regression as a post-hoc, learned model as a point of comparison for the other saliency techniques discussed. We train a single model where the input is the frequency of the dataset's vocabulary that is present in the a sentence (if a word is present in the sentence, frequency for that word is non zero, words in the vocabulary that are not in the current sentence have frequency 0). Every word in the dataset's vocabulary is included except for words that either occur less than 5 times in the entire dataset or are present in more than 80% of the sentences. This has the effect of removing words that contain somewhat nonsensical weights or stop words respectively. The labels are 0 (negative) and 1 (positive) as is typical in sentiment classification tasks. The loss function is binary cross entropy loss, and the model runs for 1000 iterations over the entire dataset. Because there is now a single weight for (almost) every word, saliency maps are determined by iterating through the tokens of a sentence and finding that tokens corresponding weight.

| **Evaluation specifics** | **Type** |
|---|---|
| **Task** | Sentiment Analysis |
| **Data** | SST-2, Yelp, Movie Rationales |
| **Evaluation Type** | Functionally grounded |
| **Application Type** | Prototype |
| **Type of deployment** | Lab-based |
| **Amount of risk** | Little-No risk |

Table 1. Evaluation Specifics

| **Technique** | **Explanation Category 1** | **Explanation Category 2** | **Explanation Category 3** |
|---|---|---|---|
| **Attention** | Learned Technique | Local Explanation | Non-Gradient based |
| **Logistic Regression** | Learned Technique | Global Explanation | Non-Gradient based |
| **LIME** | Post-hoc Technique | Local Explanation | Non-Gradient based |
| **LIG** | Post-hoc Technique | Local Explanation | Gradient based |
| **LxA** | Post-hoc Technique | Local Explanation | Gradient based |

Table 2. Categories that chosen XAI techniques fall into

# E MISCELLANEOUS ALIGNMENT SCORES

## E.1 RQ1

In this section, we list pairwise alignments between the techniques we analyzed in RQ1.

| Problem Type | Metric | Attention | LR | LIME | LIG | LxA |
|---|---|---|---|---|---|---|
| Alignment with **Attention** | **IoU** | - | 0.32 | 0.32 | 0.35 | 0.33 |
| | **RP** | - | 50.61% | 56.46% | 56.72% | 54.48% |
| Alignment with **LR** | **IoU** | 0.32 | - | 0.20 | 0.46 | 0.39 |
| | **RP** | 50.61% | - | 40.67% | 72.06% | 64.69% |
| Alignment with **LIME** | **IoU** | 0.32 | 0.20 | - | 0.26 | 0.30 |
| | **RP** | 56.46% | 40.67% | - | 45.21% | 49.68% |
| Alignment with **LIG** | **IoU** | 0.35 | 0.46 | 0.26 | - | 0.52 |
| | **RP** | 56.72% | 72.06% | 45.21% | - | 69.43% |
| Alignment with **LxA** | **IoU** | 0.33 | 0.39 | 0.30 | 0.52 | - |
| | **RP** | 54.48% | 64.69% | 49.68% | 69.43% | - |

Table 3. Pairwise alignment scores b/w XAI techniques

## E.2 RQ2

In this section, we list alignments of the tokens extracted using SST-2 + MR with the reference tokens provided by our reference for RQ1. We believe this is not a meaningful score since the alignment is w.r.t tokens calculated with SST-2 + Yelp but nevertheless, we provide these scores below.

| Alignment Type | Metric | Attention | LR | LIME | LIG | LxA |
|---|---|---|---|---|---|---|
| Alignment with references from RQ1 | **IoU** | 0.0105 | 0.0395 | 0.0149 | 0.0347 | 0.0396 |
| | **RP** | 12.33% | 38.66% | 20.33% | 30.33% | 35.33% |

Table 4. Alignment of RQ1's references with candidate spurious tokens extracted from SST+MR setup

# F EXAMPLES OF CANDIDATE SPURIOUS TOKENS

## F.1 RQ1

The tokens enumerated here are as per the setup defined in the main paper where RQ1 is solved w.r.t SST-2 and Yelp. Additionally, we show the **first 300** spurious tokens that we find. More spurious tokens can be seen in the Excel file attached.

*F.1.1* ***Reference Tokens:*** superficial, depressed, redundant, succeeds, crude, oblivious, sharply, recycled, proves, vivid, pointless, touching, emptiness, woven, cruel, historically, wisdom, fluid, seal, casts, masterpiece, hopeless, clumsy, conviction, realistic, portrayal, reveals, uneven, unpredictable, epic, considerable, enables, flashes, vicious, jagged, slick, essentially, haunting, pale, affecting, intelligent, backward, thrill, powerful, torture, celebrates, strongest, irrelevant, canvas, smug, captures, offensive, tale, sitcom, abuse, suffers, drown, enduring, acid, meaningful, juvenile, portrait, compassion, imitation, bleak, convincing, affection, kills, courage, meditation, universal, awkwardly, detached, clever, faithful, exaggerated, capable, daring, emerge, deadly, striking, dramatically, longest, skill, fierce, fallen, drain, remarkably, nausea, brightly, playful, tremendous, acute, absorbing, qualities, gripping, startling, fatal, glossy, rises, achievement, graceful, frantic, amusing, stirring, builds, flourish, touched, problematic, drains, magnetic, monstrous, breathe, asleep, borrowed, creates, expresses, inactive, establishes, feast, lavish, reduces, sympathetic, clarity, chuckle, commanding, fabricated, uneasy, precisely, stalls, jerking, harmless, adequately, eerie, insulting, racist, confident, hollow, sensation, sincerity, sensual, cerebral, glued, reflective, snaps, viewed, credible, combines, corpse, wins, lifts, shorter, refuses, elaborate, bravery, spontaneous, quietly, portrays, profound, dubious, hampered, niche, slack, vital, sweetly, infectious, conspicuous, thud, represents, exploitation, triumph, spends, breakthrough, unmistakable, capture, penetrating, fascination, economical, costly, tuned, displays, padded, propaganda, directorial, repeated, misery, willingness, pity, wears, assurance, impact, sly, emerges, static, photographed, arbitrary, riot, loosely, honesty, dreadful, elephant, approaches, straining, translate, lowered, disgust, slip, stylized, fires, utilizing, pollution, achieving, profile, intensely, loser, gently, stretched, honorable, symbolic, freedom, weary, whale, prejudice, lifeless, compatible, embraced, denial, banged, capturing, collectively, prolonged, rejected, flatly, phenomenon, emphasizes, scarcely, expressive, successfully, bitten, transformed, downward, evolve, dies, examines, heroic, foolish, formed, offended, fused, reaches, effectively, precise, playfully, bolt, screams, knockout, rooted, strains, fiercely, wreckage, ordeal, plateau, discarded, humming, commands, prominent, yields, stakes, muted, naive, worm, vacant, grenade, transforms, liability, slapping, motionless, numb, overview, vain, satisfactory, scoring, explosive, disastrous, stripped, attracting, effectiveness, counts, ineffective, error, deficit, arrogant, beaches, destructive, contributions, flopped, weighs, shaking, respects, ensures

*F.1.2* ***Logistic Regression:*** parental, the, wit, only, and, characters, something, beautiful, human, utterly, remain, filmmakers, up, could, superficial, too, merit, patriot, director, an, demonstrates, of, games, as, suicidal, depressed, thought, are, in, through, absurd, do, for, where, how, bad, greatest, intelligence, redundant, succeeds, casting, actress, face, by, ways, feminist, conspiracy, anything, black, performances, brave, enriched, spirits, or, cinema, plot, find, weaving, clone, funny, storyline, real, crude, silly, heroes, oblivious, existence, sharply, point, nowhere, recycled, already, been, fourth, swords, proves, off, has, horrors, vivid, splashed, any, pointless, beautifully, situations, room, touching, legend, less, veins, rich, apparent, joy, ugly, video, emptiness, relentless, exploration, sour, our, pretty, these, mergers, woven, aerial, storylines, experience, cruel, women, deception, eastwood, earnest, pays, homage, laughs, few, endure, help, power, generates, highs, without, historically, significant, tribute, actors, generate, comedy, heat, robert, extremes, continuum, manages, intensity, family, flair, keep, animated, alternating, parody, android, intriguing, comic, pulp, difference,

artworks, does, commonplace, explore, criminal, directed, washington, certainly, long, images, eyes, tear, cinematic, addition, titles, generally, down, mystery, 3000, own, em, other, concert, authority, fluid, had, storytelling, casts, uses, pleasures, masterpiece, called, lot, back, forgotten, laugh, car, get, hopeless, charming, dialogue, clumsy, sentiment, tissues, betrayed, surprisingly, makeup, bourne, conviction, brings, proper, role, diplomacy, generosity, impressions, artist, engaging, similar, collapse, flaws, poses, tool, excuse, effective, community, tradition, conquer, set, phones, delivers, roll, suspense, consistent, lacking, portrayal, realistic, trying, tries, stunts, reveals, talents, important, service, da, relief, filmmaking, promise, view, wait, auschwitz, simulation, theatrical, camp, extraordinary, heart, champion, adaptation, interesting, descends, works, keeping, derivative, sequel, archives, despite, argues, ambitions, problem, compelling, mere, 84, procession, remarkable, date, short, pleasant, slightest, wildly, videos, music, color, preliminary, style, narrative, difficult, unpredictable, running, spectacle, conclusions, dramas, reach, satisfying, ponder, forces, speak, enables, considerable, large, flat, bold, effects, whole, flashes, sibling, reconciliation, warmth, avoids, script, credits, bar, enjoy, predecessors, visions, beyond, pacing

*F.1.3* ***Attention:*** what, on, notorious, or, to, past, compared, this, other, for, road, taken, have, getting, know, me, but, been, because, like, of, re, has, service, one, with, advantage, account, about, that, doing, being, years, experience, do, feel, i, going, too, morning, an, get, in, same, had, it, problem, so, summer, seem, time, put, long, a, next, them, wait, you, else, order, kitchen, main, obvious, food, taking, some, tables, initial, miss, hit, seems, quality, n, than, when, (, something, stores, car, much, kind, clothes, trip, find, still, piano, picture, looking, crowd, take, from, while, hear, your, musical, found, bite, more, mixed, grab, eat, spears, play, listen, ?, our, hour, maximum, at, populated, note, anyone, ask, if, idea, making, needed, around, night, busy, first, actually, b, see, wedding, side, only, fine, opened, tickets, give, heck, down, used, groups, cheaper, 3d, into, once, since, decent, stock, any, noble, somehow, /, exact, packed, half, ve, peak, south, want, inside, where, popularity, chicken, river, practice, knew, meal, bore, asian, us, stars, thought, rice, by, look, few, table, desire, yet, out, eye, cause, amount, dragon, hole, end, think, simply, would, carry, club, photo, done, m, right, moved, fact, how, type, describe, kids, family, different, hundred, ice, 4, lights, got, :, milk, street, ahead, corners, somewhat, flavor, cutting, lower, already, buy, need, help, seen, thing, say, probably, card, competition, gift, light, bright, head, early, cut, college, part, day, school, desperately, american, friends, typical, 1998, might, generally, friend, reason, her, cell, dogs, possible, most, sleep, humane, afford, pain, could, sign, decided, another, someone, procedure, read, may, working, times, young, work, thinking, happen, conversation, near, party, others, sports, praises, drive, stop, his, enough, little, bit, those, year, totally, white, otherwise, sweet, item, meat, less, store, giant, remember, remarkable, interior, together, sisters, close, served, bloody, group, destination, reminded, slightly, feels, throughout

*F.1.4* ***LIME:.*** new, units, hide, contains, something, human, rather, nature, beautiful, characters, loves, throughout, remains, remain, satisfied, could, worst, far, treatment, merit, film, still, small, games, director, turn, personal, hollywood, emotional, poetry, thought, films, deeply, goes, like, used, make, anymore, movies, part, happening, saw, bad, movie, story, greatest, musicians, cold, intelligence, usual, concept, face, young, whose, actress, projects, swimming, casting, even, original, ways, see, black, named, anything, storm, dick, dirty, camps, karen, conspiracy, smile, comes, performances, brave, cast, mixed, spirits, half, worse, world, cinema, good, alternative, nothing, start, finish, plot, action, little, often, find, interest, acted, poorly, year, sit, another, best, theme, funny, think, real, issues, marriage, liked, adults, tucked, storyline, silly, heroes, existence, sharply, every, point, course, entire, sense, dog, classic, nowhere, sometimes, dry, come, times, already, care, count, fourth, feature, territory, covers, suggesting, game, million, version, 40, gorgeous, cross, lost, performance, touch, bringing, us, ever, given, beauty, bloody, many, situations, poor, ben,

driving, directions, room, touching, starts, legend, less, rich, stuff, veins, joy, apparent, video, shot, digital, ugly, core, ending, exploration, though, hold, ford, pretty, damned, never, feel, one, current, industry, recording, climate, corporate, circus, together, experience, overall, shots, aerial, men, women, total, carried, lack, sounds, ideas, seem, fresh, history, weak, people, power, help, almost, horror, absolute, life, much, without, fun, based, TRUE, significant, historically, tribute, first, two, enough, heat, appear, comedy, reaction, throw, actors, ca, generate, robert, family, keep, intensity, manages, entertainment, animated, central, written, believe, anyone, creation, around, questions, difference, comic, moments, become, standard, follow, criminal, explore, yet, sudden, wisdom, career, directed, washington, clear, ahead, certainly, memory, long, eyes, away, read, images, hate, tear, addition, recent, titles, sporting, generally, quite, go, going, show, late, put, science, theatre, build, website, theater, mystery, em, work, concert, deal, direction, authority, harris, ron, seal, fluid, would, gone, step, piece, harsh, shirt, takes

*F.1.5* ***LIG:.*** parental, new, contains, wit, characters, nature, and, beautiful, rather, throughout, utterly, remain, filmmakers, up, worst, superficial, tragic, too, patriot, emotional, demonstrates, film, small, director, personal, with, out, suicidal, depressed, deeply, most, absurd, goes, t, happening, bad, story, lend, dumb, greatest, intelligence, his, usual, redundant, concept, succeeds, casting, face, projects, above, young, original, ways, feminist, conspiracy, theorist, smile, brave, performances, enriched, spirits, cast, worse, world, cinema, alternative, viewing, plot, cylinders, poorly, find, year, sit, weaving, funny, theme, another, clone, storyline, crude, real, issues, heroes, existence, sharply, nowhere, point, course, entire, dog, recycled, count, been, care, feature, fourth, covers, territory, version, million, gorgeous, swords, performance, proves, touch, superb, horrors, vivid, beauty, any, pointless, beautifully, situations, room, touching, subtle, starts, less, sophisticated, veins, rich, joy, shot, ugly, exploration, emptiness, relentless, core, thrilling, hold, ford, damned, mergers, climate, recording, current, storylines, woven, experience, together, empathy, cruel, deception, sounds, lack, carried, seem, history, eastwood, earnest, homage, few, laughs, power, endure, help, horror, generates, life, historically, significant, TRUE, tribute, first, two, enough, throw, spark, robert, extremes, continuum, intensity, manages, flair, terrific, family, entertainment, animated, believe, anyone, central, parody, intriguing, android, pulp, comic, difference, alternating, questions, moments, artworks, stale, yet, wisdom, washington, respectable, showcase, eyes, images, read, tear, hate, cinematic, recent, addition, generally, mystery, put, 3000, em, late, go, work, fluid, direction, deal, ron, storytelling, piece, casts, landscape, attractive, takes, pleasures, masterpiece, time, lot, hopeless, charming, college, pure, match, creations, dialogue, sentiment, clumsy, unfortunately, tissues, bring, betrayed, conviction, bourne, jason, role, brings, diplomacy, charm, generosity, years, share, effort, engaging, collapse, forgive, flaws, tool, serves, learning, effective, community, importance, conquer, business, delivers, plans, cell, roll, suspense, surprise, lacking, portrayal, realistic, grab, trying, lump, stunts, reveals, talents, others, relief, welcome, filmmaking, promise, wait, auschwitz, theatrical, simulation, death, faith, extraordinary, champion, heart, adaptation, descends, noir, tight, works, into, derivative, sequel

*F.1.6* ***LxA:.*** parental, new, from, only, nature, characters, about, beautiful, something, and, same, remain, to, filmmakers, could, up, worst, superficial, treatment, tragic, too, s, patriot, demonstrates, film, can, ., still, small, personal, games, director, with, as, of, out, a, suicidal, poetry, thought, through, most, deeply, films, ‘, than, in, are, more, absurd, goes, lengths, t, do, make, those, part, was, movie, some, dignity, story, greatest, musicians, cold, intelligence, his, usual, redundant, actress, succeeds, whose, above, face, young, is, by, original, even, ways, conspiracy, feminist, theorist, named, black, anything, dirty, your, smile, comes, brave, enriched, spirits, cast, mixed, which, worse, half, or, cinema, world, alternative, good, viewing, start, but, finish, plot, cylinders, little, find, acted, often, interest, will, year, weaving, clone, best, funny, another, storyline, real, between, have, silly, i, crude,

think, heroes, existence, sharply, course, classic, dog, point, entire, dry, times, having, recycled, count, care, territory, covers, game, version, million, gorgeous, swords, them, proves, performance, off, he, lost, bringing, touch, once, has, again, superb, horrors, vivid, any, us, beauty, bloody, many, beautifully, poor, him, driving, subtle, touching, be, legend, starts, sophisticated, veins, rich, joy, shot, ugly, digital, exploration, relentless, emptiness, at, ending, just, though, hold, feel, never, one, mergers, recording, current, climate, industry, storylines, woven, experience, magnificent, shots, cruel, deception, carried, empathy, men, women, total, marginal, fresh, history, earnest, pays, homage, weak, skip, few, surprises, people, endure, generates, life, fun, historically, TRUE, appear, spark, first, two, throw, actors, reaction, so, extremes, continuum, manages, intensity, flair, terrific, family, entertainment, animated, anyone, you, written, central, alternating, parody, pulp, questions, difference, moments, artworks, become, explore, standard, stale, !, sudden, wisdom, ahead, memory, showcase, tear, long, eyes, images, hate, cinematic, addition, recent, titles, going, em, mystery, other, late, down, go, concert, work, fluid, ron, direction, cam, further, would, had, storytelling, piece, casts, landscape, watch, talented, takes

### F.2 RQ2

The tokens for SST-2 + Yelp are as listed above. In contrast, some tokens for the in-domain case of SST-2 + MR are listed below.

*F.2.1* ***LIME:.*** one, two, even, get, head, since, see, life, go, church, party, find, lost, idea, bad, deal, makes, films, movie, generally, drive, highway, break, watch, attempt, generation, continues, write, guys, drink, cool, accident, plot, couples, harder, girlfriend, mess, package, presents, dies, teen, time, like, back, know, still, another, going, little, got, within, power, across, really, story, william, brother, start, russian, action, ship, comes, lee, happy, crew, empty, quick, damn, starring, regarding, bringing, feels, kick, flash, tech, virus, substance, sequences, three, name, much, however, music, make, based, show, late, things, car, television, hair, course, police, immediately, wrong, evidence, nice, seem, tells, gets, sounds, squad, stuff, movies, tale, cuts, first, year, best, last, came, century, production, dead, reason, empire, cast, pretty, recent, piece, beat, 20th, contest, worried, warner, animation, mouse, quest, steal, cartoon, promising, may, people, part, way, film, man, number, love, due, mother, special, yet, fact, type, boy, industry, takes, produce, falls, pictures, category, gain, string, favor, stable, comments, kills, theatrical, potentially, seemingly, new, many, called, john, face, something, 13, director, title, taking, lot, whole, especially, fight, previous, writer, fighting, apparently, otherwise, planet, ideas, scenes, horror, accused, assault, mistake, expert, believes, horrible, world, order, making, business, short, position, stop, eight, ask, surface, becoming, influence, murder, beneath, sight, values, sick, twisted, necessarily, surveillance, long, hand, felt, thing, months, nine, eye, talking, pass, exactly, acting, grant, jim, huge, adam, nervous, hugh, terrible, smiles, August, built, road, career, far, european, call, hours, families, nearly, hour, plays, actor, drama, difficult, presence, walking, screen, unable, trip, swedish, express, spend, depth, relationships, significance, secrets, accurate, reflection, adds, years, well, used, eyes, old, young, french, rest, killed, mean, parents, okay, ways, anyway, evil, tim, plain, sees, revenge, fourteen, work, want, every, given, big, role, works, playing, michael, usually, performance, character, dr, bit, studio, kiss, scottish, brian, opposite, remembered, check, hidden, occasionally, substantial, brings, brilliant, halfway, cox, also, good, along, almost, political, running, fall, campaign, couple, ago, truth, flat, dogs, jay, romantic, cats, manages, made, team, american, version, france, popular, style, van, prior, double, goes, asian, asia, kong, hong

*F.2.2* ***Logistic Regression:*** with, his, that, has, is, and, in, of, for, but, one, it, an, to, on, this, cool, the, break, continues, guys, into, go, even, drive, her, teen, mess, makes, package, lost, which, get, idea, touches, films, what, since, party, him, bad, they, life, presents, then, plot, such, attempt, movie,

see, deal, girlfriend, out, two, watch, head, generally, your, very, generation, find, all, bringing, still, bug, empty, feels, action, russian, damn, kick, time, little, power, really, quick, no, another, within, comes, happy, substance, there, story, when, going, lee, start, was, like, starring, do, know, got, here, across, back, why, sequences, we, few, not, as, by, music, are, cute, wrong, tale, invention, nice, show, things, so, seem, same, much, up, tells, does, car, make, stuff, course, television, however, name, under, movies, three, late, more, these, based, gets, hair, sounds, viewer, police, after, from, be, century, quest, colorful, piece, cast, 20th, last, beat, recent, lively, flawed, dead, contest, had, pretty, came, first, best, production, if, cartoon, other, its, animation, year, only, their, promising, reason, he, film, love, may, string, number, industry, fatal, part, mother, produce, theatrical, falls, seemingly, boy, favor, man, pictures, endless, potentially, yet, takes, re, due, fact, type, people, way, category, both, special, lot, ideas, himself, assault, expert, called, writer, something, horrible, scenes, mistake, whole, title, homage, fighting, horror, previous, 13, taking, off, otherwise, planet, face, especially, fight, john, new, director, many, who, those, beneath, loses, becoming, wo, itself, sick, world, just, or, short, ask, sight, stop, order, making, twisted, how, values, yourself, about, surface, business, murder, can, hugh, smiles, were, talking, grant, hand, terrible, me, annoying, laughs, acting, long, thing, exactly, pass, jim, felt, huge, adam, eye, far, too, accurate, actor, trip, difficult, relationships, secrets, built, nearly, reflection, express, spend, presence, painfully, european, depth, drama, plays, loser, hours, screen, sentimental, road, adds, until, hour, career, call, ways, dude, french, killed, parents, old, years, plain, well, some, you, eyes, rest, revenge, mean, where, evil, boring, before, young, used, sees, should, halfway, indie, brings, playing, bit, substantial, check, big, usually, works, want, brilliant, michael, occasionally, performance, thriller, studio, kiss, role, flick, work, every, character, given, through, remembered, romantic, almost, good, running, along, ago, dogs, truth, manages, couple, political, flat, also, fall, satire, surprises, at, have, popular, france, stunt, double, bland, been, down, version, kong, joke, attempts, style, hong, prior, made, team, now, cinematic, american, filmmaking, goes, vehicle, provide, levels, entire

*F.2.3* ***Attention:*** out, into, two, what, then, such, your, even, get, head, since, see, life, go, very, church, party, find, lost, idea, bad, deal, makes, films, movie, generally, drive, highway, review, break, watch, attempt, generation, continues, write, guys, drink, cool, accident, plot, couples, harder, girlfriend, mess, package, presents, ), t, was, all, when, there, time, no, we, like, back, do, know, still, another, here, going, little, few, got, within, why, power, across, really, story, william, brother, start, russian, action, ship, comes, lee, happy, crew, empty, quick, damn, starring, regarding, bringing, feels, kick, flash, as, by, are, not, up, after, so, more, three, under, these, same, name, much, however, music, make, based, show, late, things, car, does, television, hair, course, police, immediately, wrong, evidence, nice, seem, tells, gets, sounds, squad, stuff, movies, tale, cuts, from, had, be, first, their, its, other, if, only, year, best, last, came, century, production, dead, reason, empire, cast, pretty, recent, piece, beat, 20th, contest, worried, animation, he, may, people, part, both, way, re, film, man, number, love, due, mother, special, yet, fact, type, boy, industry, takes, produce, falls, pictures, category, gain, string, favor, comments, 18, 4, new, many, off, called, john, face, something, himself, 13, director, title, taking, lot, whole, especially, fight, previous, writer, fighting, apparently, otherwise, planet, ideas, scenes, horror, accused, assault, mistake, expert, believes, 8, or, who, about, can, just, world, how, those, order, making, business, short, position, stop, eight, itself, ask, surface, becoming, influence, murder, beneath, sight, yourself, values, sick, twisted, /, were, me, long, hand, felt, thing, months, eye, talking, pass, exactly, acting, grant, jim, huge, adam, nervous, hugh, terrible, until, too, built, road, career, far, european, call, hours, families, nearly, hour, plays, actor, drama, difficult, presence, walking, screen, unable, trip, swedish, express, spend, depth, relationships, significance, secrets, !, you, some, where, before, years, well, used, eyes, old, should, young, french, rest, killed, mean, parents, okay, ways, anyway, evil, tim, plain, sees, revenge, through, work, want, every, given, big,

role, works, playing, michael, usually, performance, character, dr, bit, studio, kiss, scottish, brian, opposite, remembered, check, hidden, occasionally, substantial, brings, o, also, good, along, almost, political, running, fall, couple, ago, truth, flat, dogs, jay, romantic, at, have, been, made, now, down, team, american, version, france, popular, style, van, prior, goes, asian, asia, kong, hong, vehicle, attempts, *, did, must, moment, entire, provide, multiple, levels, necessary, fair, clearly, she, game, together, woman, la, living, either, stone, question, image, dream, reality, anne, biggest, dreams, l, will, went, york, live, always, case, looking, someone, science, looks, escape, fiction, vampires, latest, suppose, m, most, second, large, sure, green, earth, saying, novel, guy, starting, stars, beautiful, minor, wonder

*F.2.4* ***LIG:.*** watch, drive, generation, generally, girlfriend, head, go, harder, films, cool, life, movie, party, one, bad, deal, accident, lost, teen, makes, no, t, why, was, all, few, when, here, ), there, do, we, start, got, feels, tech, comes, sequences, power, within, time, going, still, ship, story, like, another, happy, bringing, flash, quick, brother, really, russian, lee, not, same, so, up, under, by, are, as, more, these, does, after, sounds, cute, cuts, seem, tale, nice, late, music, name, wrong, course, movies, viewer, tells, gets, things, squad, make, stuff, only, from, its, their, if, be, had, other, year, colorful, quest, recent, best, cast, lively, 20th, piece, first, dead, century, flawed, cartoon, beat, animation, reason, re, both, he, produce, pictures, film, falls, love, favor, endless, part, way, due, yet, off, 4, himself, 18, title, carpenter, lot, planet, something, fighting, writer, previous, director, supposedly, horror, mistake, assault, homage, fight, expert, new, or, how, yourself, just, can, who, those, itself, about, 8, beneath, ask, twisted, short, world, making, murder, surface, eight, were, me, /, hugh, nervous, eye, grant, hand, smiles, thing, huge, talking, exactly, long, too, until, road, depth, secrets, hour, drama, swedish, families, sentimental, screen, difficult, built, actor, plays, painfully, reflection, hours, accurate, nearly, !, should, you, before, where, some, mean, used, evil, french, killed, rest, tim, well, years, ways, old, revenge, through, hannibal, thriller, dr, works, bit, usually, halfway, studio, work, performance, occasionally, big, substantial, brings, check, brilliant, michael, kiss, o, manages, political, satire, couple, good, romantic, ago, also, been, down, have, at, now, version, kong, hong, prior, asia, style, cinematic, made, popular, classics, double, france, team, american, did, *, ironic, must, provide, multiple, entire, imagery, levels, moment, she, question, either, game, together, dream, image, biggest, stone, will, l, case, live, always, went, someone, latest, most, m, novel, adaptation, wonder, sure, critic, crush, starting, minor, beautiful, horrific, predictable, earth, green, second, than, d, third, space, would, results, thousand, odyssey, hell, hope, everything, whatever, considering, successful, left, times, ultimately, ve, mine, familiar, likes, say, trailer, teenage, journey, heard, pleasing, worth, maybe, definitely, any, my, during

*F.2.5* ***LxA:.*** they, into, go, and, one, with, attempt, but, i, then, find, lost, accident, him, out, even, to, ., his, on, cool, s, very, has, is, life, mess, movie, for, ?, idea, party, of, films, guys, (, two, church, get, break, kick, going, really, damn, back, flash, happy, within, substance, bringing, comes, all, sequences, start, we, got, tech, still, another, t, few, was, lee, know, little, empty, like, why, tells, things, immediately, cute, as, car, tale, up, music, are, not, so, by, name, seem, under, based, however, more, course, wrong, 20th, their, recent, colorful, century, from, year, lively, best, beat, production, cast, last, steal, be, quest, dead, only, piece, mother, part, due, re, endless, people, stable, love, man, string, favor, special, seemingly, both, falls, industry, he, 13, writer, title, off, whole, 4, something, fight, supposedly, homage, new, especially, previous, fighting, himself, lot, many, expert, face, horrible, carpenter, taking, about, world, twisted, or, beneath, yourself, sick, sight, making, stop, business, those, just, position, 8, desires, long, me, jim, /, months, hugh, eye, huge, grant, were, adam, swedish, reflection, call, mundane, express, significance, families, presence, plays, depth, hours, accurate, drama, trip, some, years, evil, tim, ways, well, eyes, french, !, rest, dude, you, brings, bit, through, work, performance, works, occasionally, michael, halfway, hidden, hannibal,

big, along, satire, fall, political, ago, also, almost, manages, dogs, good, surprises, campaign, have, cinematic, vehicle, down, france, double, made, kong, team, version, attempts, necessary, fair, *, did, moment, provide, game, la, reality, nightmare, question, woman, either, she, dream, stone, live, looking, science, someone, large, green, most, strikes, adaptation, minor, guy, beautiful, choosing, m, crush, earth, predictable, novel, said, successful, times, hope, space, hell, third, close, than, would, various, elements, everything, kind, sent, odyssey, mix, left, nothing, ultimately, pleasing, mine, familiar, leads, journey, joan, crazy, shows, law, heard, trailer, actually, during, scene, grand, pair, any, wow, sell, minutes, tailor, features, monster, darkness, gun, them, lots, violence, machine, chick, showing, men, behind, tired, creepy, brutally, lies, killer, source, upon, girls, potential, sort, executive, already, cheap, found, worst, disappointing, gives, could, becomes, worse, hardly, glow, targets, day, set, group, attention, middle, daniel, around, dancing, full, historical, highly, experience, perspective, intelligence, member, wave, passion, phone, taken, america, cell, tough, town, invented, easily, coherent, never, 1, numerous, derivative, twists, turns, increasingly, pile, games, screenplay, entertaining, greatest, outrageous, studios, disney, addition, possible, future, race

### F.3 RQ3

| LR x Evidences | Sentence count | Class Skew |
|---|---|---|
| mother | 34 | 0.71 |
| horrible | 15 | 0.20 |
| trip | 15 | 0.27 |
| green | 14 | 0.71 |
| predictable | 19 | 0.26 |
| put | 41 | 0.27 |
| knows | 19 | 0.79 |
| human | 35 | 0.71 |
| fails | 14 | 0.21 |
| liked | 11 | 0.82 |
| fast | 20 | 0.25 |
| wonderful | 27 | 0.70 |
| slow | 21 | 0.19 |
| decent | 22 | 0.27 |
| impossible | 17 | 0.29 |
| welcome | 14 | 0.71 |
| sit | 16 | 0.13 |
| musical | 13 | 0.77 |
| heavy | 14 | 0.14 |
| moving | 11 | 0.73 |
| forced | 15 | 0.73 |
| enjoyed | 11 | 0.73 |
| central | 12 | 0.83 |
| tradition | 11 | 0.73 |
| loved | 14 | 0.71 |
| intelligent | 19 | 0.74 |
| spielberg | 15 | 0.80 |
| force | 18 | 0.78 |
| adult | 13 | 0.77 |
| stunning | 14 | 0.86 |

*F.3.1* ***Spurious tokens.*** LR x Evidences contains the list of tokens present in both human highlights provided in the Movie Rationales dataset and our spurious correlations list obtained from using Logistic regression. We threshold for tokens present in at least 10 sentences, and provide the sentence count and percentage of sentences for which the model predicted a positive sentiment. The tables for the spurious correlations found using the other techniques can be found in the spreadsheet.

**Human Annotations Example 1:**

yet another brainless teen flick \nthis one is about , surprise , drugs and sex .\nstars katie holmes and sarah\npolly could n\'t look more bored \ntheir characters are cardboard cut - outs of every cliched teenager out there .\none thing you need to know\nis i really hated this movie .\neverything about it annoyed the hell out of me .\nthe acting , and script , the plot , and ending .\nthe director ( of the fluke hit swingers ) could have very well directed a bunch of no - name actors and had a watchabe film .\nthe " big " stars of go pretty much drown the project of any originality .\ni felt like i was watching dawson \'s creek episode 200 .\nalthough the film still would have stayed at red despite its cast .\nthe " surprise " ending was sooo predictable .\nsince when is a male character \'s sudden outing of the closet considered a surprise in hollywood anymore ? ?\ngo is dawson \'s creek + varsity blues - she \'s all that\n= go home and watch something else .

**Human Annotations Example 2:**

"peter jackson 's the frighteners has received some notice for setting the record for most computer effects ever in a movie , and still coming in at the extremely cheap $ 30 million price tag .\nbut for those who were dismayed by this year 's blockbusters like twister and independence day , the frighteners has much more to offer than special effects .\nand for those worried wether or not peter jackson would compromise to hollywood you can rest easily .\nthe frighteners is as far removed from hollywood as a high - profile movie can get .\nmichael j. fox stars as frank bannister , a con artist who can speak to ghosts .\nhe uses this ability to set up a scam in a small town where his ghost buddies scare the hell out of people , then he comes and pretends to get rid of them .\nthis is how he has made a living ever since his wife died in a car crash 5 years ago .\nfrank 's latest customers are a young couple , played by trini alvarado and peter dobson .\nwhen dobson ends up dead , alvarado starts to take an interest in fox .\nbut dobson 's spirit is still around as he refuses to beleive\nhe 's dead .\nthis leads to a very awkward and amusing dinner date between fox and alvarado , with dobson tagging along as a ghost .\nthings start getting complicated for fox when he is accused for a series of murders taking place in the town .\nfox sees someone named the soul collector crushing the heart of the victims , but noone else can see that .\nso when fox shows up to try and save each victim , naturally people suspect he is the killer .\nfox sees that alvarado is next on the soul collector 's hit list , and the last half hour of the movie deals with fox 's attempts to save her from this evil spirit .\nthere are many wonderful twists and turns in the screenplay written by peter jackson and frances walsh .\nthe movie starts off as a black comedy , and ends up a horror - action film .\nthe mix between these genres are perfect .\nno laughs are sacrificed in the name of horror , and vice versa .\none point of contention might be a lackluster score by danny elfman .\nbut that hardly seems like a flaw when you have such a diverse cast all in top form .\nmichael\nj. fox delivers one of his best performances to date as a man who hides the sorrow of his wife 's death , and then is forced to confront this later on .\nalvarado , looking like andie macdowell , makes a great frightened , tough , and smart heroine .\nand jeffery combs , as a paranoid fbi agent , is brilliantly bizaare .\nthe frighteners never once feels like it is running long .\nthe first hour is as funny as any comedy this year , and the last half hour is as thrilling as any of the big budget blockbusters .\nthis movie is probably what casper would 've looked like if david lynch directed it .\nit 's easily the best film of the year , so far ."

*F.3.2* ***Human evidence***. Human highlights obtained from the Movie Rationales dataset