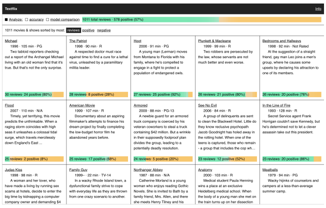
≡ CONTENTS

Cloudera Fast Forward

Textflix: Using Transfer Learning for a Natural Language Processing Prototype



Textflix

Info

■ Analyze: □ accuracy □ model comparison 1011 total reviews · 578 positive (57%)

1011 movies & shows sorted by most: reviews positive negative

Michael 1996 · 105 min · PG Two tabloid reporters checking out a report of the Archangel Michael living with an old woman find that it's true. But that's not the only surprise. 30 reviews· 24 positive (80%)	The Patriot 1998 · 90 min · R A respected doctor must race against time to find a cure for a lethal virus, unleashed by a paramilitary militia leader. 28 reviews· 8 positive (28%)	Hoot 2006 · 91 min · PG A young man (Lerman) moves from Montana to Florida with his family, where he's compelled to engage in a fight to protect a population of endangered owls. 27 reviews· 25 positive (92%)	Plunkett & Macleane 1999 · 99 min · R Two robbers are persecuted by the law, whose servants are not much better and even worse. 26 reviews· 21 positive (80%)	Bedrooms and Hallways 1998 · 92 min · Not Rated At the suggestion of a straight friend, gay man Leo joins a men's group, where he causes some upsets by declaring his attraction to one of its members. 26 reviews· 20 positive (76%)
Flood 2007 · 110 min · N/A Timely, yet terrifying, this movie predicts the unthinkable. When a raging storm coincides with high seas it unleashes a colossal tidal surge, which travels mercilessly down England's East ... 25 reviews· 2 positive (8%)	American Movie 1999 · 107 min · R Documentary about an aspiring filmmaker's attempts to finance his dream project by finally completing the low-budget horror film he abandoned years before. 25 reviews· 17 positive (68%)	Armored 2009 · 88 min · PG-13 A newbie guard for an armored truck company is coerced by his veteran coworkers to steal a truck containing $42 million. But a wrinkle in their supposedly foolproof plan divides the group, leading to a potentially deadly resolution. 24 reviews· 5 positive (20%)	See No Evil 2006 · 84 min · R A group of delinquents are sent to clean the Blackwell Hotel. Little do they know reclusive psychopath Jacob Goodnight has holed away in the rotting hotel. When one of the teens is captured, those who remain - a group that includes the cop wh… 23 reviews· 12 positive (52%)	In the Line of Fire 1993 · 128 min · R Secret Service agent Frank Horrigan couldn't save Kennedy, but he's determined not to let a clever assassin take out this president. 23 reviews· 20 positive (86%)
Judas Kiss 1998 · 98 min · R A woman and her lover, who have made a living by running sex scams at hotels, decide to enter the big time by kidnapping a computer company owner and demanding $4	Family Guy 1999– · 22 min · TV-14 In a wacky Rhode Island town, a dysfunctional family strive to cope with everyday life as they are thrown from one crazy scenario to another.	Northanger Abbey 1987 · 88 min · N/A Catherine Morland is a young woman who enjoys reading Gothic Novels. She is invited to Bath by a family friend, Mrs. Allen, and there she meets Henry Tilney and his	Anatomy 2000 · 103 min · R Medical student Paula Henning wins a place at an exclusive Heidelberg medical school. When the body of a young man she met on the train turns up on her dissection	Meatballs 1979 · 94 min · PG Wacky hijinks of counselors and campers at a less-than-average summer camp.

The Textflix prototype

Perhaps the most exciting aspect of transfer learning is that it is so broadly applicable. It does not enable one specific

≡ CONTENTS

capability, like question answering, language translation, text classification, or search. It makes *each* of these more accurate, at a lower cost, with less data, and thus more accessible to a broader community. It opens the pathway from research to production.

We built a prototype, Textflix, that leverages transfer learning for sentiment analysis. Textflix performs positive or negative sentiment detection on movie reviews, which present a challenge to text processing systems because of the complex ways in which humans express their preferences. We added LIME, an interpretability technique, to provide insight into the model's predictions and used this mechanism to provide individual summaries of each movie. The entire product was

≡ CONTENTS

built using a model trained on just 500 labeled examples. The modeling was implemented with off-the-shelf tools from AllenNLP and did not require writing any code. Everything was completed within an infrastructure budget of $25.

Sentiment

While transfer learning is likely to be beneficial for almost any NLP application, some will benefit more than others.

Even though transfer learning can be implemented at low cost, we do not recommend using a sophisticated deep learning model for a 1% gain in accuracy. Some applications, like basic topic categorization, can be handled admirably by extremely simple statistical models (like Naive Bayes, which is essentially counting words). In such cases, sticking with the simple model works best.

But sentiment detection – the task of inferring how the writer or speaker feels – requires sophisticated tools. When humans convey their feelings, they may use sarcasm and irony. They might incorporate obscure pop culture references. They might have mixed feelings, contradict themselves, or even change their minds by the time they've finished their thoughts! Without transfer learning, all of these difficult challenges would need to be solved by learning from a single, possibly small, dataset. Because of these difficulties, sentiment detection demonstrates very clearly the power of transfer learning.

Dataset

≡ CONTENTS

Textflix is built on the IMDB dataset, a popular open dataset commonly used for sentiment analysis. The dataset consists of 25,000 movie reviews from the users of the popular online movie database, although only 500 of those were used to build the model behind Textflix, as a demonstration of the power of transfer learning.

While sentiment analysis can be applied in many different domains, movie reviews are interesting because they present a diverse set of challenges. Many of the reviews are straightforward and simple, but some contain subtle clues as to the author’s opinion. These subtleties are challenging for machine learning models – a model based on simple statistics will not work well. Because the reviews are written in rather plain English (no specialized dialects or slang), publicly available pretrained models (which were trained on generic English) will work well.

Models

The modeling process for Textflix was extremely simple. One of the great benefits of transfer learning is that it eliminates the need to invent complex new neural network architectures that uniquely solve a particular problem. Off-the-shelf transfer learning models already provide state-of-the-art accuracy, so the fewer changes we make, the better. After comparing several models, we ended up using the large version of the BERT model…

We experimented with several popular transfer learning models, and also compared their performance to simple but strong baseline methods.

≡ CONTENTS

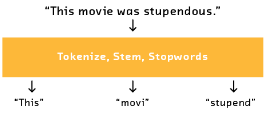
Although we built Textflix with a model trained on only 500 examples, for each model we explored its performance curve when training on more labeled examples. In the following sections we present these results and justify our modeling choices.

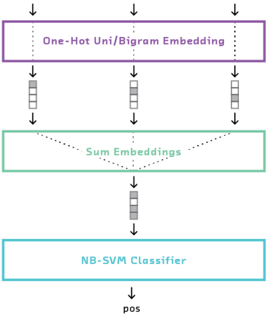
Baseline Models

Testing baseline methods is an important first step in the modeling process. Even if they are unlikely to yield a usable model, they are easy to implement, present a logical reference for comparing future models to, and are usually more interpretable than more sophisticated alternatives. In some cases, the simplicity and interpretability advantages they present may outweigh the decrease in accuracy in the results they produce. For our prototype, we explored two baselines: SVM with Naive Bayes features and word vectors.

NB-SVM

For text classification problems like sentiment analysis it makes sense to choose a simple model based on bag-of-words as the first baseline. In many text classification problems, like topic classification, these types of baseline models may even be the best choice. NB-SVM treats the text as a bag of words and combines a Naive Bayes model (also a reasonable baseline) with a support vector machine. This model has been shown to produce strong linear baselines for text classification, and sentiment analysis in particular.



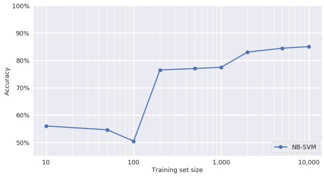


Pipeline architecture for the NB-SVM model.

In testing this model we generated both uni- and bigram features for the NB-SVM classifier, removed stopwords from the input, and used a

Snowball stemmer to normalize each word.

The performance of the NB-SVM was poor and unpredictable at low dataset sizes – it simply did not have enough observations to learn which words strongly correlated with positive or negative sentiment. However, at larger training set sizes (e.g., 10,000 examples), this baseline reached a useful accuracy reached a useful accuracy of about 85% with very little tuning.



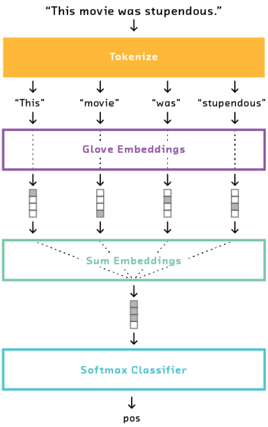
NB-SVM is no better than guessing at training set sizes less than 200.

This performance curve shows that even simple bag-of-words models can identify a large majority of sentiment examples correctly. And because the NB-SVM model is simple and fast to implement, it would be easier to support in a production use case than some deep neural network models.

Word Vectors

As we've seen, the NB-SVM model cannot be reasonably expected to perform well with small training datasets. Because it is a bag-of-words model, it has no a priori it has no a priori knowledge of words and their associations and must learn everything from the training data. A model that leverages word vectors should do better in the small-data regime, since word vectors are a form of transfer learning. That is, the meaning of words is already captured in the pretrained word vectors, and is not affected by the small data size.

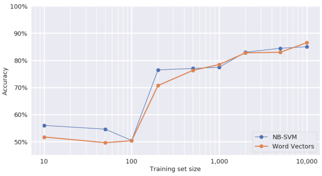
We used the following simple architecture for this model.



Pipeline architecture for

`the word vectors model.`

The performance curve shows that the word vectors model generally follows the same trend as the NB-SVM baseline. (It is important to note that there is a lot of variance at each data point, which is not shown in the plot, so small differences should be taken lightly.)



`Transferred word vectors suffer from the same problems as the NB-SVM model.`

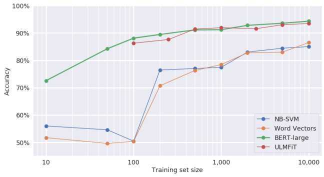
This result is, in some ways, surprising since word vectors are a form of transfer learning and should therefore be more resilient to limited data than the baseline NB-SVM model. It is likely that the word vectors model could be tuned to outperform NB-SVM at smaller training sizes by adjusting the hyperparameters and architecture. However, we purposely spent little time optimizing hyperparameters with any of the models. The hyperparameter optimization step is largely a heuristic process and can require deep expertise to guide the search – something we wanted to avoid.

Transfer Learning Models

With well-performing baselines established, we began to try out several of the publicly available transfer learning models. We experimented with ULMFiT using the Fast.AI library and found it to perform well, even with limited data. We performed language-model fine-tuning using 50,000 unlabeled examples for the ULMFiT

≡ CONTENTS

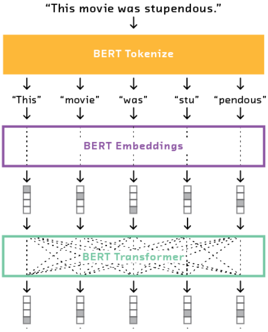
model, and then performed supervised training at various dataset sizes. In addition to ULMFiT, we ran experiments for both the BERT-Base and BERT-Large models, using the AllenNLP library.

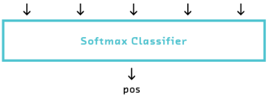


New transfer learning models can perform well even with fewer than 100 examples.

The best-performing models were the BERT-Large and ULMFiT models. While these models produced nearly equal results, we found that the BERT-Large model was easier to implement and experiment with. This is in part because no language model fine-tuning step was required with BERT-Large (target task fine-tuning, however, was tuning, however, was performed for each model), and in part because ULMFiT is trained via the Fast.AI library, which imposes development via notebooks – something we find to be restrictive.

We trained the BERT-Large model without making any custom modifications and defined the entire experiment in a JSON configuration file in the AllenNLP style. We trained on a single GPU for 20 epochs, using an Adam optimizer, and used gradual unfreezing for fine-tuning the layers of the model.





Pipeline architecture for the BERT-large model

Overall, we did very little tuning of the model’s hyperparameters. This limited tuning requirement is one of the greatest benefits of transfer learning: the out-of-the-box performance is already very good, and eking out a further 1-2% of accuracy has diminishing returns and would require expensive-to-acquire knowledge of the model architecture.

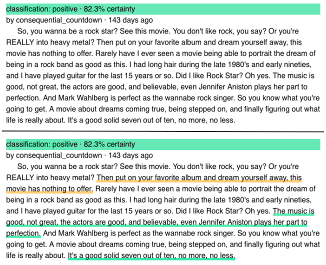
The final BERT model provides accuracy roughly equivalent to the state-of-the-art model, using just 500 labeled examples for training.

Interpretability

A cost of using transfer

≡ CONTENTS

A cost of using transfer learning is the less interpretable, black box nature of neural network models. The ability to detect sentiment is undeniably useful, but the ability to explain those predictions is significantly more powerful. With interpretability, we can not only classify the user's sentiment but point to specific evidence to support that classification. We found the addition of interpretability to the model's predictions to be surprisingly useful. Here we will discuss our approach and some of the benefits.



classification: positive · 82.3% certainty

by consequential_countdown · 143 days ago

So, you wanna be a rock star? See this movie. You don't like rock, you say? Or you're REALLY into heavy metal? Then put on your favorite album and dream yourself away, this movie has nothing to offer. Rarely have I ever seen a movie being able to portrait the dream of being in a rock band as good as this. I had long hair during the late 1980's and early nineties, and I have played guitar for the last 15 years or so. Did I like Rock Star? Oh yes. The music is good, not great, the actors are good, and believable, even Jennifer Aniston plays her part to perfection. And Mark Wahlberg is perfect as the wannabe rock singer. So you know what you're going to get. A movie about dreams coming true, being stepped on, and finally figuring out what life is really about. It's a good solid seven out of ten, no more, no less.

classification: positive · 82.3% certainty

by consequential_countdown · 143 days ago

So, you wanna be a rock star? See this movie. You don't like rock, you say? Or you're REALLY into heavy metal? Then put on your favorite album and dream yourself away, this movie has nothing to offer. Rarely have I ever seen a movie being able to portrait the dream of being in a rock band as good as this. I had long hair during the late 1980's and early nineties, and I have played guitar for the last 15 years or so. Did I like Rock Star? Oh yes. The music is good, not great, the actors are good, and believable, even Jennifer Aniston plays her part to perfection. And Mark Wahlberg is perfect as the wannabe rock singer. So you know what you're going to get. A movie about dreams coming true, being stepped on, and finally figuring out what life is really about. It's a good solid seven out of ten, no more, no less.

Predicting sentiment is useful, but pointing readers to specific evidence that the model

≡ CONTENTS

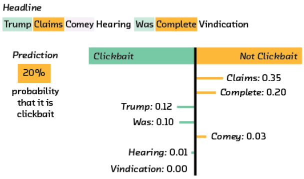
relies on allows the reader to trust the prediction.

Modified LIME for Sequences

We used the technique of Local Interpretable Model-agnostic Explanations (LIME) to add interpretability. LIME can be applied to almost any type of model, but here we will consider its application to text data. Put simply, LIME is a way to understand how different parts of an input affect the output of a model. This is accomplished, essentially, by turning the dials of the input and observing the effect on the output.

Consider a model that classifies news headlines as “clickbait” or “not clickbait.” The model could identify individual words as more or less clickbaity, depending on how they affect the output.

If removing a word from the input generally makes the model predict clickbait, then that word contributes to the model's clickbait classification for that particular input.

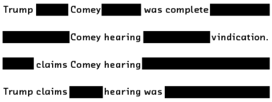


LIME identifies words that contribute to positive or negative classification.

This method makes sense for models that process the input as a bag of words, but what about for models like BERT that process the input as a sequence? These models don't view an input as simply a collection of isolated words; they are capable of picking out statements, sentences, or thoughts. Further, removing words from the input as a method to understand the model's

≡ CONTENTS

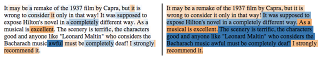
understand the model's predictions makes the input incoherent.


Trump [redacted] Comey [redacted] was complete [redacted]
[redacted] Comey hearing [redacted] vindication.
[redacted] claims Comey hearing [redacted]
Trump claims [redacted] hearing was [redacted]

LIME defaults to masking words, which produces a type of incoherent language that sequence models like BERT were not trained on.

What works well for a bag-of-words model seems unnatural for a sequence model. One could imagine that BERT works instead by treating input text as compositions of whole sentences, rather than compositions of words (although this is a simplification). In this case, it makes more sense to apply LIME at the sentence level. Removing sentences may still produce incoherent text, but if thoughts are relatively confined to sentences then this may be an acceptable trade-off.

For these reasons, we applied LIME at the sentence level for the final BERT model. For the NB-SVM baseline we explored both word- and sentence-level LIME.



It may be a remake of the 1937 film by Capra, but it is wrong to consider it only in that way! It was supposed to expose Hilton's novel in a completely different way. As a musical is excellent. The scenery is terrific, the characters good and anyone like "Leonard Maltin" who considers the Bacharach music awful must be completely deaf! I strongly recommend it.

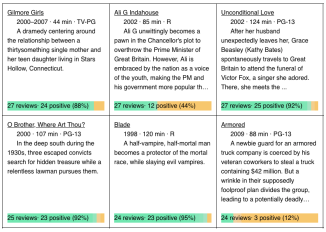
It may be a remake of the 1937 film by Capra, but it is wrong to consider it only in that way! It was supposed to expose Hilton's novel in a completely different way. As a musical is excellent. The scenery is terrific, the characters good and anyone like "Leonard Maltin" who considers the Bacharach music awful must be completely deaf! I strongly recommend it.

Left: Word-level interpretability can be overwhelming and difficult to parse. Right: Sentence-level interpretability provides a more concise picture.

Regardless of the model, we found sentence-level LIME easier to digest. When highlighting individual words there is simply too much information to process.

Interpretability Provides Trust

With a model that can gauge the sentiment of a movie review, we can do some useful things. For some useful things. For example, we can visualize the overall sentiment of a particular film. This allows users to quickly gauge popular movies and could be used as a tool for recommending the best movies.



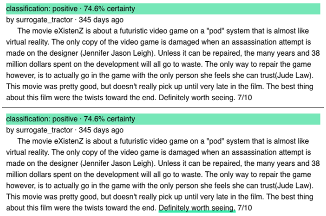
Gilmore Girls 2000–2007 · 44 min · TV-PG A dramedy centering around the relationship between a thirtysomething single mother and her teen daughter living in Stars Hollow, Connecticut. 27 reviews· 24 positive (88%)	Ali G Indahouse 2002 · 85 min · R Ali G unwittingly becomes a pawn in the Chancellor's plot to overthrow the Prime Minister of Great Britain. However, Ali is embraced by the nation as a voice of the youth, making the PM and his government more popular th… 27 reviews· 12 positive (44%)	Unconditional Love 2002 · 124 min · PG-13 After her husband unexpectedly leaves her, Grace Beasley (Kathy Bates) spontaneously travels to Great Britain to attend the funeral of Victor Fox, a singer she adored. There, she meets the ... 27 reviews· 25 positive (92%)
O Brother, Where Art Thou? 2000 · 107 min · PG-13 In the deep south during the 1930s, three escaped convicts search for hidden treasure while a relentless lawman pursues them. 25 reviews· 23 positive (92%)	Blade 1998 · 120 min · R A half-vampire, half-mortal man becomes a protector of the mortal race, while slaying evil vampires. 24 reviews· 23 positive (95%)	Armored 2009 · 88 min · PG-13 A newbie guard for an armored truck company is coerced by his veteran coworkers to steal a truck containing $42 million. But a wrinkle in their supposedly foolproof plan divides the group, leading to a potentially deadly… 24 reviews· 3 positive (12%)

The Textflix dashboard makes it easy to find popular and unpopular movies.

Users can drill down to a particular title and read a sampling of its positive and negative reviews. But model predictions are rarely blindly accepted. Users who want to verify that the model is doing something sensible may opt to skim the reviews, looking for evidence of the model's predictions. This introduces a significant

≡ CONTENTS

cognitive burden – reviews are long, but the sentiment may be expressed in just one sentence. The following review, where most of the text is simply summarizing the plot, is a good example.



classification: positive · 74.6% certainty
by surrogate_tractor · 345 days ago

The movie eXistenZ is about a futuristic video game on a "pod" system that is almost like virtual reality. The only copy of the video game is damaged when an assassination attempt is made on the designer (Jennifer Jason Leigh). Unless it can be repaired, the many years and 38 million dollars spent on the development will all go to waste. The only way to repair the game however, is to actually go in the game with the only person she feels she can trust(Jude Law). This movie was pretty good, but doesn't really pick up until very late in the film. The best thing about this film were the twists toward the end. Definitely worth seeing. 7/10

classification: positive · 74.6% certainty
by surrogate_tractor · 345 days ago

The movie eXistenZ is about a futuristic video game on a "pod" system that is almost like virtual reality. The only copy of the video game is damaged when an assassination attempt is made on the designer (Jennifer Jason Leigh). Unless it can be repaired, the many years and 38 million dollars spent on the development will all go to waste. The only way to repair the game however, is to actually go in the game with the only person she feels she can trust(Jude Law). This movie was pretty good, but doesn't really pick up until very late in the film. The best thing about this film were the twists toward the end. Definitely worth seeing. 7/10

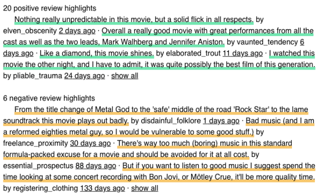
Top: The model classifies this review as positive, but with no supporting evidence. Bottom: The model explains its positive classification by highlighting a single sentence, which should be enough to convince users to trust the prediction. View the review at on Texflix

Without interpretability, a user may need to skim each and every review for evidence supporting the model's predictions. But using LIME, we can automatically point to specific evidence that the model relies on to make its predictions. In this case, users can immediately see why the model has made a prediction and trust that it is correct.

Interpretability as a Summarization Tool

Interpretability tools that can point to specific evidence in the input to explain model predictions

actually offer a new product opportunity. Specifically, we found that using LIME to identify highly polarized sentences allowed us to construct high-level summaries of each movie’s reviews.



20 positive review highlights

Nothing really unpredictable in this movie, but a solid flick in all respects. by elven_obscenity 2 days ago · Overall a really good movie with great performances from all the cast as well as the two leads, Mark Walhberg and Jennifer Aniston. by vaunted_tendency 6 days ago · Like a diamond, this movie shines. by elaborated_trout 11 days ago · I watched this movie the other night, and I have to admit, it was quite possibly the best film of this generation. by pliable_trauma 24 days ago · show all

6 negative review highlights

From the title change of Metal God to the 'safe' middle of the road 'Rock Star' to the lame soundtrack this movie plays out badly. by disdainful_folklore 1 days ago · Bad music (and I am a reformed eighties metal guy, so I would be vulnerable to some good stuff.) by freelance_proximity 30 days ago · There's way too much (boring) music in this standard formula-packed excuse for a movie and should be avoided for it at all cost. by essential_prospectus 88 days ago · But if you want to listen to good music I suggest spend the time looking at some concert recording with Bon Jovi, or Mötley Crue, it'll be more quality time. by registering_clothing 133 days ago · show all

≡ CONTENTS

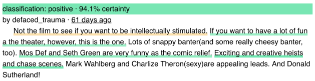
By scoring each sentence as positive or negative, interpretability tools enable summarizations of sentiment for each movie.

These summaries give a reader a broad picture of others' sentiment toward the movie, without them having to scroll through each of the individual reviews. Without interpretability, we could only show users entire reviews, which are burdensome to read. We find interpretability is an essential component of building interesting data products.

Successes and Failures

Transfer learning using BERT yielded impressive results, especially considering the limited training data. It was a clear and significant improvement over the alternatives. But like any model, BERT has its model, BERT has its limitations. In this section we examine some of the concrete successes and failures of the prototype model.

BERT Identifies Mixed Feelings



classification: positive · 94.1% certainty
by defaced_trauma · 61 days ago
Not the film to see if you want to be intellectually stimulated. If you want to have a lot of fun a the theater, however, this is the one. Lots of snappy banter(and some really cheesy banter, too). Mos Def and Seth Green are very funny as the comic relief. Exciting and creative heists and chase scenes. Mark Wahlberg and Charlize Theron(sexy)are appealing leads. And Donald Sutherland!

Interpretability at the sentence level shows how the model balances opposing sentiments within reviews. View the review on Textflix

There are many reviews in the dataset that would classify as “easy” in the sense that they communicate clear sentiment using words that are positive or negative. The final deployed BERT model, unsurprisingly, handles these cases well. The passage above demonstrates some of this plain language, but is

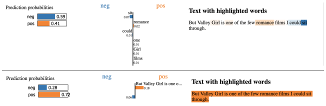
≡ CONTENTS

interesting in that it has individual sentences that communicate opposing sentiments. With the added interpretability mechanism, it is possible to not only show the model’s overall prediction for the sentiment of the review but also its predictions for some of the individual parts. This helps draw attention to reviews that are mixed, where a single sentiment score does not necessarily tell the entire story.

BERT Handles Subtle Language and Negation Well

> “But Valley Girl is one of the few romance films I could sit through.”
> – A review for *Valley Girl*

This review text is relatively subtle. It does not contain any explicitly positive or negative words, yet it is fairly obvious to a human that it conveys a positive sentiment.



Text with highlighted words

But Valley Girl is one of the few romance films I could sit through.

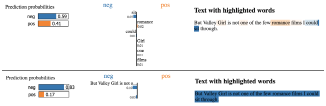
Text with highlighted words

But Valley Girl is one of the few romance films I could sit through.

Top: BERT correctly predicts positive sentiment despite the indirect language. Bottom: NB-SVM incorrectly predicts negative sentiment. Neutral words like “romance” and “sit” are associated with positive or negative feelings.

The BERT model produces the correct prediction, even if it is not clear how it infers this. The baseline NB-SVM model, on the other hand, must invoke simple rules which fail in this case. Using word-level LIME, we can see that the baseline model associates neutral words like “romance” and “sit” with positive or negative sentiment, even though these words on their own do not convey sentiment.

As a test of BERT's understanding, we can experiment with negating the entire phrase and observing the new prediction.



Text with highlighted words

But Valley Girl is not one of the few romance films I could sit through.

Text with highlighted words

But Valley Girl is not one of the few romance films I could sit through.

`Top: BERT correctly flips its prediction when the “not” modifier is used to negate the positive statement. Bottom: NB-SVM does not recognize “not” as a negative word, and does not change its prediction.`

Indeed, BERT now predicts a negative sentiment. The baseline model, however, can only rely on each word as a small piece of evidence. Even though it sees “not” as carrying negative sentiment value, the prediction overall is still positive.

≡ CONTENTS

There are cases where BERT fails, though, even with simple negation. Although LIME is a useful tool, it’s not a catch-all. There is still quite some mystery in terms of how BERT works and when it will fail.

Sentiment Can Be Unclear

> “In any event, a very, very fun, but fairly bad, movie.”
> – A positive review for *See No Evil*

> “Average adventure movie that took a serious story and ‘Holywoodised’ it… The screenplay was average. The charm of Connery made up for his wrong Arabic accent and all the scenes with President T. Roosevelt were masterpiece takes.”

> – A positive review for *The Wind and the Lion*

> “The storyline is okay at best, and the acting is surprisingly alright, but after awhile it’s gets to be a little much. But, still it’s fun, quirky, strange, and original.”
> – A positive review for *Don’t Look in the Basement*

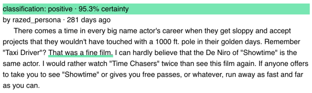
Each of these examples is from an instance where the BERT model made an incorrect prediction. Each of them is surprising, even to humans, in some way. That is, they each express a statement of sentiment that is opposite of the rating the reviewer gave. This underscores the fact that sentiment analysis is not really a binary task – in reality, humans may like some things and dislike others, which makes it somewhat nonsensical to distill their feelings into a

≡ CONTENTS

single positive or negative classification. In some cases where the model gets it wrong, a human might also.

One encouraging aspect of these cases is that the BERT model is often quite uncertain in its predictions. Models will inevitably make errors, but the ability to provide a reliable measure of uncertainty makes it easier to use and trust their outputs.

Entity Assignment Presents a Challenge



classification: positive · 95.3% certainty
by razed_persona · 281 days ago
There comes a time in every big name actor's career when they get sloppy and accept projects that they wouldn't have touched with a 1000 ft. pole in their golden days. Remember "Taxi Driver"? That was a fine film. I can hardly believe that the De Niro of "Showtime" is the same actor. I would rather watch "Time Chasers" twice than see this film again. If anyone offers to take you to see "Showtime" or gives you free passes, or whatever, run away as fast and far as you can.

A review for *Showtime* where the model is confused by a positive reference to another film.

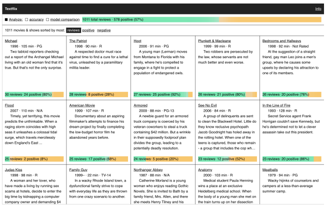
This is a review for the film *Showtime*, a comedic cop film featuring Robert

De Niro and Eddie Murphy. The review is quite negative, but features a mention of a separate movie - *Taxi Driver* - which the reviewer remarks is a “fine film.” The BERT model does not recognize that this positive sentiment is actually attached to a different entity and should therefore be disregarded.

While BERT has been shown to be effective at the task of entity recognition, the fine-tuned model here was never explicitly taught to do this. In fact, since the model only has access to the review itself, it really has no way to know that *Taxi Driver* is not the subject of this review. However, even if the model were given the movie title associated with each review, it is unlikely that it would learn such nuances. This is because it would not have nearly enough examples in its small training set for it to be able to learn this special case. Even including more examples might not be sufficient – instead, the training objective might need to be changed to explicitly ask the model to pick out entities and assign each of them a sentiment score.

This is an interesting example because it shows that even though transfer learning models may have the skills to perform particular tasks, they only invoke those skills if it is important for prediction accuracy during training. Transfer learning models are powerful, but they still fall far short of human intuition.

Product Design



Textflix Info

■ Analyze: □ accuracy □ model comparison 1011 total reviews · 578 positive (57%)

1011 movies & shows sorted by most: reviews positive negative

Michael 1996 · 105 min · PG Two tabloid reporters checking out a report of the Archangel Michael living with an old woman find that it's true. But that's not the only surprise. 30 reviews· 24 positive (80%)	The Patriot 1998 · 90 min · R A respected doctor must race against time to find a cure for a lethal virus, unleashed by a paramilitary militia leader. 28 reviews· 8 positive (28%)	Hoot 2006 · 91 min · PG A young man (Lerman) moves from Montana to Florida with his family, where he's compelled to engage in a fight to protect a population of endangered owls. 27 reviews· 25 positive (92%)	Plunkett & Macleane 1999 · 99 min · R Two robbers are persecuted by the law, whose servants are not much better and even worse. 26 reviews· 21 positive (80%)	Bedrooms and Hallways 1998 · 92 min · Not Rated At the suggestion of a straight friend, gay man Leo joins a men's group, where he causes some upsets by declaring his attraction to one of its members. 26 reviews· 20 positive (76%)
Flood 2007 · 110 min · N/A Timely, yet terrifying, this movie predicts the unthinkable. When a raging storm coincides with high seas it unleashes a colossal tidal surge, which travels mercilessly down England's East ... 25 reviews· 2 positive (8%)	American Movie 1999 · 107 min · R Documentary about an aspiring filmmaker's attempts to finance his dream project by finally completing the low-budget horror film he abandoned years before. 25 reviews· 17 positive (68%)	Armored 2009 · 88 min · PG-13 A newbie guard for an armored truck company is coerced by his veteran coworkers to steal a truck containing $42 million. But a wrinkle in their supposedly foolproof plan divides the group, leading to a potentially deadly resolution. 24 reviews· 5 positive (20%)	See No Evil 2006 · 84 min · R A group of delinquents are sent to clean the Blackwell Hotel. Little do they know reclusive psychopath Jacob Goodnight has holed away in the rotting hotel. When one of the teens is captured, those who remain - a group that includes the cop wh… 23 reviews· 12 positive (52%)	In the Line of Fire 1993 · 128 min · R Secret Service agent Frank Horrigan couldn't save Kennedy, but he's determined not to let a clever assassin take out this president. 23 reviews· 20 positive (86%)
Judas Kiss 1998 · 98 min · R A woman and her lover, who have made a living by running sex scams at hotels, decide to enter the big time by kidnapping a computer company owner and demanding $4	Family Guy 1999– · 22 min · TV-14 In a wacky Rhode Island town, a dysfunctional family strive to cope with everyday life as they are thrown from one crazy scenario to another.	Northanger Abbey 1987 · 88 min · N/A Catherine Morland is a young woman who enjoys reading Gothic Novels. She is invited to Bath by a family friend, Mrs. Allen, and there she meets Henry Tilney and his	Anatomy 2000 · 103 min · R Medical student Paula Henning wins a place at an exclusive Heidelberg medical school. When the body of a young man she met on the train turns up on her dissection	Meatballs 1979 · 94 min · PG Wacky hijinks of counselors and campers at a less-than-average summer camp.

The final Textflix design.

We make prototypes to spark our clients' imaginations about how emerging technologies could be applied to their own business problems. Besides being functional, this means our prototypes need to tell a good story. In this section we'll discuss the design and storytelling decisions that went into Textflix.

Visualizing the Classification

Once we decided to focus on sentiment analysis and use the IMDB review dataset, the fundamental unit of the prototype was clear: the text of a movie review and the model's classification of that review as positive or

≡ CONTENTS

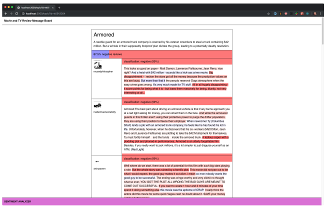
negative. Presenting just those elements makes for a functional tech demo, but also a dry one.

The first step we took towards making things more interesting *and* understandable was to apply the LIME interpretability technique at the sentence level. Using LIME, we show the user which sentences were driving the classification. This makes the classification feel much more dynamic. You can visualize (a simplified version of) the model's examination process.

Developing the Story

The second thing we did was develop a story about why these reviews needed to be analyzed. On the dataset side, we grouped reviews under the movie or show they were reviewing, and selected the groups with the most reviews. Grouping reviews made it possible to build the prototype as an imaginary movie/show review site.

The IMDB dataset, by design, covers a wide range of movies and shows, so sorting by the most reviewed gave us a rather eccentric list of entertainments. Part of the storytelling challenge of these prototypes is explaining dataset limitations like this one in a non-disruptive way.



Movie and TV Review Message Board

Armored

A newbie guard for an armored truck company is coerced by his veteran coworkers to steal a truck containing $42 million. But a wrinkle in their supposedly foolproof plan divides the group, leading to a potentially deadly resolution.

87.5% negative reviews

rousedphilosopher

classification: negative (99%)

This looks so good on paper - Matt Damon, Lawrence Fishbourne, Jean Reno, nice right? And a heist with $42 million - sounds like a kick-ass crime movie. Big disappointment - I reckon the stars got all the money because the production values on this are lousy. But more than that it the pseudo reservoir Dogs atmosphere when the easy crime goes wrong. It's very much made for TV stuff. All in all hugely disappointing - it score points for being what it is - but loses them massively for being, bluntly, not very interesting at all..

inattentiveinevitability

classification: negative (99%)

Armored The best part about driving an armored vehicle is that if any bums approach you at a red light asking for money, you can shoot them in the face. And while the armoured guards in this thriller aren't using their protective power to purge the drifter population, they are using their position to fleece their employer. When newcomer Ty (Columbus Short) lands a job with an armored trunk company, he feels like he has found his lot in life. Unfortunately, however, when he discovers that his co- workers (Matt Dillon, Jean Reno and Laurence Fishburne) are plotting to take the $42 M shipment for themselves, Ty must fortify himself and the funds inside the armored truck. A tedious caper with a plodding plot and phoned-in performances, Armored is an utterly forgettable film. Besides, if you really want to jack millions, it's a lot simpler to just disguise yourself as an ATM. (Red Light)

shinytavern

classification: negative (99%)

Well where do we start, there was a lot of potential for this film with such big stars playing a role. But the whole story was ruined by a horrific plot. This movie did not pan out to be what i would expect, the good guy makes it out alive, i mean co mon nobody wants the good guy to be successful. The ending was cringe worthy and very cliché no thought what so ever, YOU GOT THE PLOT ALL WRONG THE BAD GUYS ARE MEANT TO COME OUT SUCCESSFUL. If you want to waste 1 hour and 2 minutes of your time spend it doing something else this movie was the epitome of CRAP. I really think the actors did this movie for some quick Vegas cash no doubt about it. SAVE your money

An earlier version of the prototype presented the reviews as part of an entertainment message board.

Originally, we thought about imagining the

≡ CONTENTS

reviews as part of a message board, which would help explain both the range of topics and the text-only nature of the reviews (movie review sites often have a mechanism for directly inputting user ratings, while message board posts are typically limited to a text box). Later we realized we didn’t have to be so specific to provide a plausible explanation for why the selection was so varied. Anyone who subscribes to a streaming service has seen the odd range of movies and TV shows that can result from behind-the-scenes license negotiations. We decided our make-believe service would be called Textflix, and that its weird selection was just a result of the licenses available.

Adding Drama to Analysis

Once we had our idea for

≡ CONTENTS

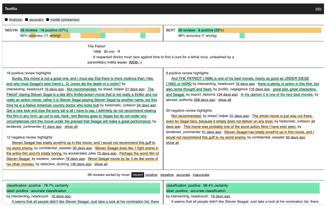
Once we had our idea for a fictional streaming site in place, we began working on showing the value sentiment analysis could bring to the site. Here again we had to balance the design of a real product with the need for storytelling and a little bit of drama. We decided that you would be able to turn the sentiment analysis on and off in the prototype (in a real product there’d be no reason for the off option). This would help us emphasize, by contrast, the capabilities text analysis gives you. Without analysis, you must read each review, one-by-one, to get an idea of the overall sentiment of the reviews. With analysis you can see the general opinion at a glance. Analysis makes sentiment computable and sortable, allowing you to answer questions like “what is the most liked movie?” (an indie drama called *What*

☰ CONTENTS

Alice Found in this case).

The minimal design of the prototype emphasizes the new powers analysis gives you. With analysis off you're faced with long blocks of text and no color. With it on, you get green and orange sentiment indicators, underlines, and review highlights.

A Peek Behind the Curtain



NBSVM 28 reviews · 16 positive (57%)
60% accuracy (11 wrong)

BERT 28 reviews · 8 positive (29%)
96% accuracy (1 wrong)

The Patriot
1998 · 90 min · R
A respected doctor must race against time to find a cure for a lethal virus, unleashed by a paramilitary militia leader. IMDB →

16 positive review highlights
Surely, this movie is not a great one, and I must say that there is more violence than I like, and why must Seagal's best friend L. Q. Jones die the death of a victim? by intersecting_headcount 16 days ago · Not recommended. by dread_helper 31 days ago · The Patriot" staring Steven Segal is a late 90's thriller/action movie that is not really a thriller and not really an action movie; rather it is Steven Segal playing Steven Segal by another name, but this time he is a Native American country doctor who kicks butt by hedonistic_collision 34 days ago · Get a new look and lose the pony tail is all I have to say. I definitely do not recommend viewing this film in any form, go out to eat, heck, rent Barney goes to Vegas but do not under any circumstances rent this movie under the precept that Seagal will make a great performance. by tenderest_commenter 41 days ago · show all

12 negative review highlights
Steven Seagal has totally scre#@d up in this movie, and I would not recommend this guff to my worst enemy. by confidential_sweater 50 days ago · Steven Seagal does like 1 fight scene in the entire film and it's totally boring. by accelerated_piles 75 days ago · Perhaps the worst film of Steven Seagal. by esoteric_variation 78 days ago · Worst Seagal movie by far (I do like some of his other movies). by defective_docking 146 days ago · show all

8 positive review highlights
And THE PATRIOT (1998) is one of his best movies, nearly as good as UNDER SIEGE (1992) or HARD by intersecting_headcount 16 days ago · there is plenty of action in this flick, but also some thought and heart. by prolific_negligence 116 days ago · great plot, great characters, and Seagal. by export_diploma 134 days ago · In my opinion it is one of his very best movies. by abroad_authority 208 days ago · show all

20 negative review highlights
Not recommended. by dread_helper 31 days ago · The whole movie is just way out there, even for Segal fans, because it simply does not deliver on any level. by hedonistic_collision 34 days ago · This movie was probably one of the worst action films I have ever seen. by tenderest_commenter 41 days ago · Steven Seagal has totally scre#@d up in this movie, and I would not recommend this guff to my worst enemy. by confidential_sweater 50 days ago · show all

28 reviews sorted by most: recent positive negative accurate inaccurate

classification: positive · 78.7% certainty
label: positive · accurate classification
by intersecting_headcount · 16 days ago
It seems that all people didn't like Steven Seagal: Just take a look at his nomination list, there

classification: positive · 98.4% certainty
label: positive · accurate classification
by intersecting_headcount · 16 days ago
It seems that all people didn't like Steven Seagal: Just take a look at his nomination list, there

A movie page with the accuracy and model comparison options turned on. These options would not be appropriate for a consumer-facing product, but we included them for their explanatory power.

With analysis on, Textflix shows the capabilities that text analysis can bring to a product. We included two further analysis options to give people a closer look at the algorithm. These would not be included in a consumer-facing product, but we put them in the prototype for their explanatory power. The first option is “accuracy.” It shows how well the model’s classifications matched the review’s original labels. With accuracy on you can see the model’s overall performance as well as find reviews where the model got things wrong. This feature is obviously useful but it is only available because we are working with an example dataset that is fully labeled, something unlikely to happen in real life (if your real life dataset is fully labeled you don’t need a model to classify it). The accuracy feature in the prototype, then, is a peek behind the curtain. It wouldn’t be possible in a

≡ **CONTENTS**

real product, but in Textflix, seeing the errors helps build your intuition about the model’s performance and where it might tend to get things wrong or right.

The second feature is “model comparison.” When activated, this feature opens a split-screen view, where our transfer-learning trained model is on the right, and NB-SVM, our baseline model, is on the left. Seeing their classifications side-by-side gives you a concrete idea of the gains made through transfer learning. You can compare the accuracy of both approaches and see reviews where they disagree. In a final consumer-facing product, you’d want to only expose the best model (though checking it against a baseline during development is still very much recommended).

CONTENTS

There were more options we could have added, but we wanted this prototype to be clearly focused on a core set of ideas. Most importantly, we wanted to show the kinds of capabilities that text analysis could open up for a product. Then we wanted to provide a view into the model's accuracy and show it in the context of a less advanced approach. We hope the final product will be both inspirational and informative for people working on similar projects.