INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

MINERÍA DE TEXTOS

ITESO, Universidad
Jesuita de Guadalajara

PROJECT #1
PROGRAMMING ASSIGNMENT: SENTIMENT ANALYSIS USING N-GRAMS

by
Sofía Maldonado García
Viviana Toledo De la Fuente
Diana Denise Valdivia Vargas

Juan Antonio Vega Fernández

Fecha: 28/09/2025

Assignment Goals

- Understand and implement n-gram models (unigram, bigram, trigram).
- Apply these models to text classification for sentiment analysis.
- Perform basic feature extraction and preprocessing.
- Evaluate model performance using appropriate metrics.
- Conduct error analysis to identify strengths and limitations of n-gram models.

Pre-Processing:

In this project we prepared a dataset of text reviews for sentiment analysis. We downloaded essential NLP resources from NLTK and built a custom tokenization function that removes line breaks, stop words, and punctuation, lemmatizes words, and tags fully uppercase words as potentially important. We applied this function to each review to generate cleaned tokens. Finally, we split the data into training, development, and test sets (50%-25%-25%) and converted the token lists back into strings so they could be used directly by our models.

```
Train dataset size: 25000 reviews
Dev dataset size: 12500 reviews
Test dataset size: 12500 reviews
```

Feature-Extraction

We used a CountVectorizer to turn our cleaned reviews into numbers. It looked at single words, two-word, and three-word phrases (n-grams). We fitted it on the training data and used the same settings on the dev and test sets. This lets us see which words and phrases appear in each review so we can use them in our models.

The purpose was to extract those n-grams from each review.

```
'act', 'act like', 'actor', 'actor film', 'actor movie', 'african', 'air', 'also'
```

Model Training

For this project, we will program three different models: Logistic Regression, Multinomial Naive Bayes and SVM across the data, and compare its metrics and confusion matrix.

Note: All of the models were trained using the same length of n-gram since they were specified on the document, but in the notebook we explore the different lengths of n-grams and their effect on the metrics.

EVALUATION

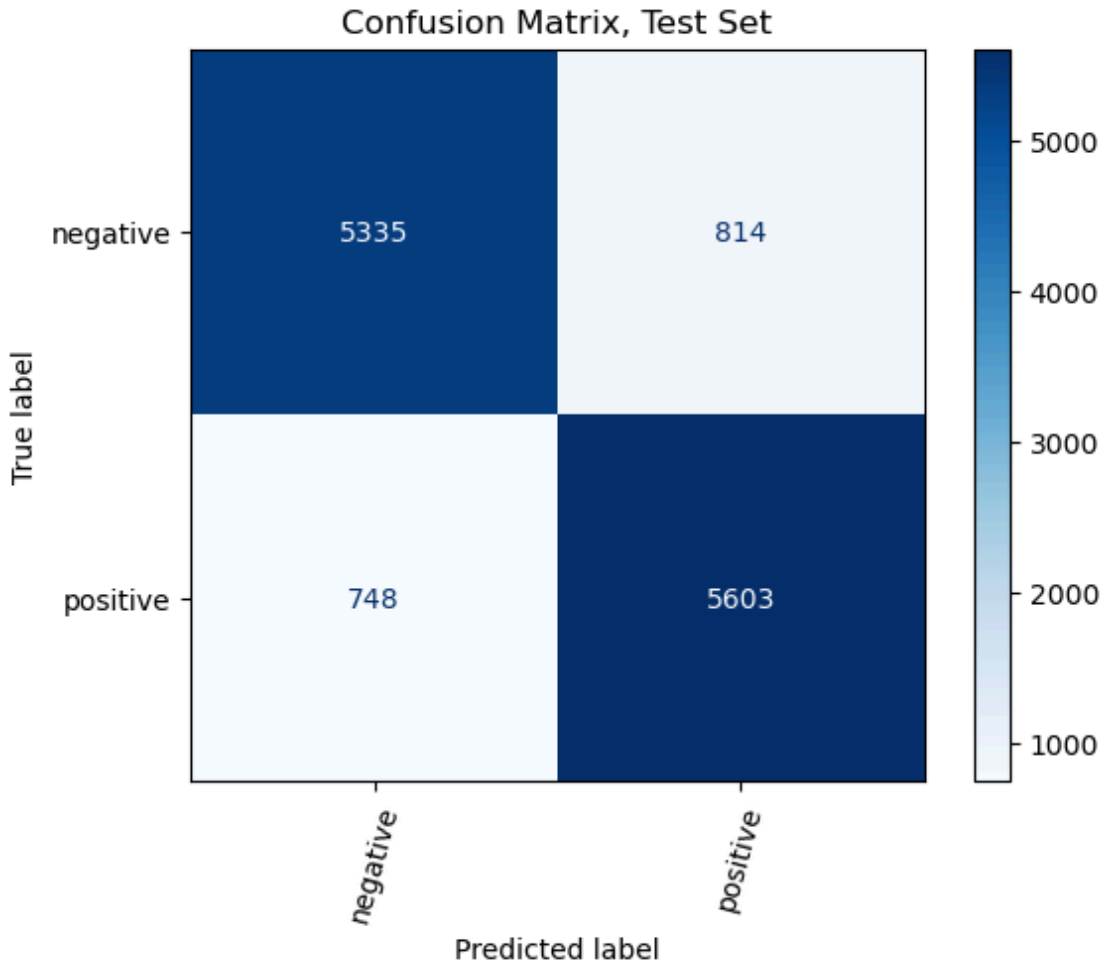
We will now present the evaluation metrics [on the test data only] of the models by using the n-gram count described on the document [1,3] and their confusion matrix:

Logistic Regression:

- **F1 Score**: 0.8750
- **Confusion Report:**

	precision	recall	f1-score	support
negative	0.8770	0.8676	0.8723	6149
positive	0.8731	0.8822	0.8777	6351
accuracy			0.8750	12500
macro avg	0.8751	0.8749	0.8750	12500
weighted avg	0.8751	0.8750	0.8750	12500

- **Confusion Matrix:**

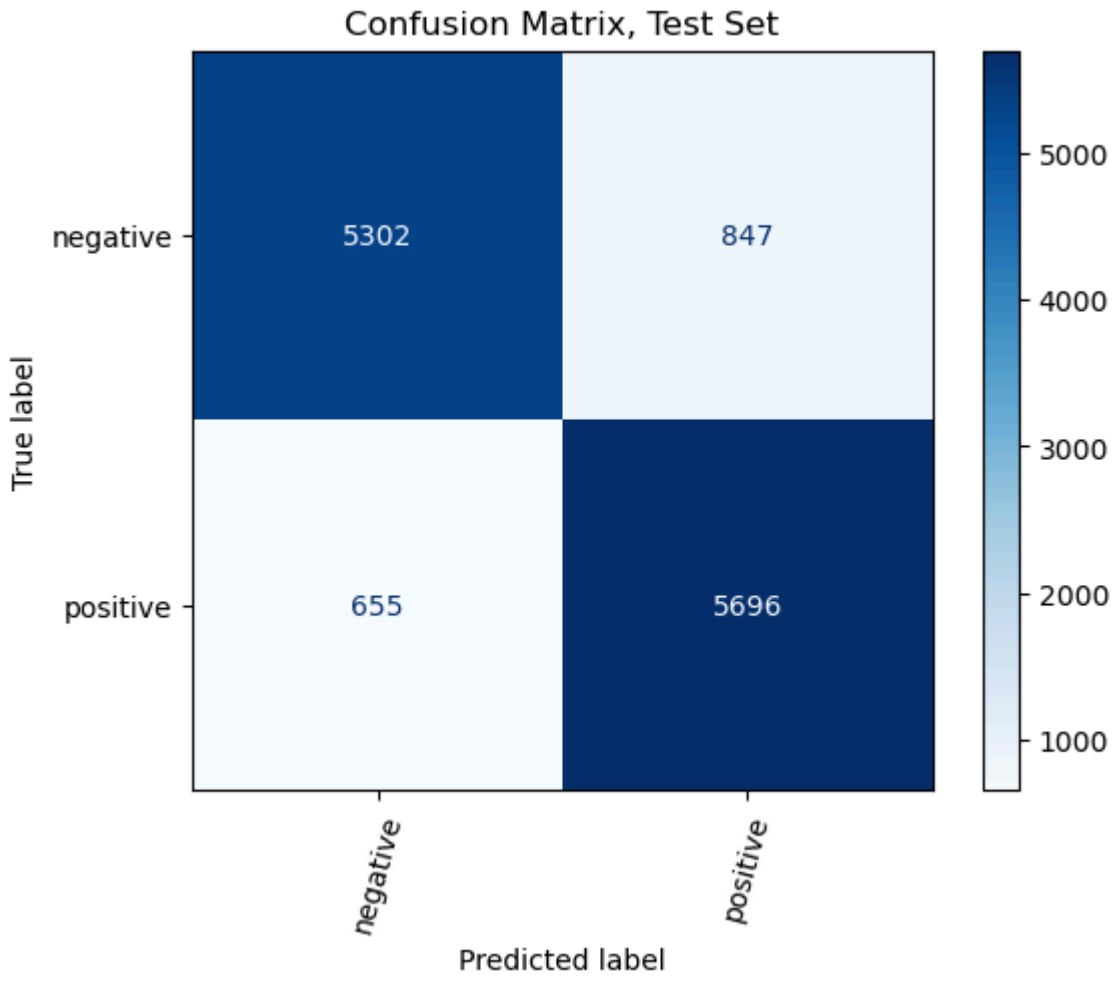


Naive Bayes:

- **F1 Score:** 0.8798
- **Confusion Report:**

	precision	recall	f1-score	support
negative	0.8900	0.8623	0.8759	6149
positive	0.8705	0.8969	0.8835	6351
accuracy			0.8798	12500
macro avg	0.8803	0.8796	0.8797	12500
weighted avg	0.8801	0.8798	0.8798	12500

- **Confusion Matrix:**

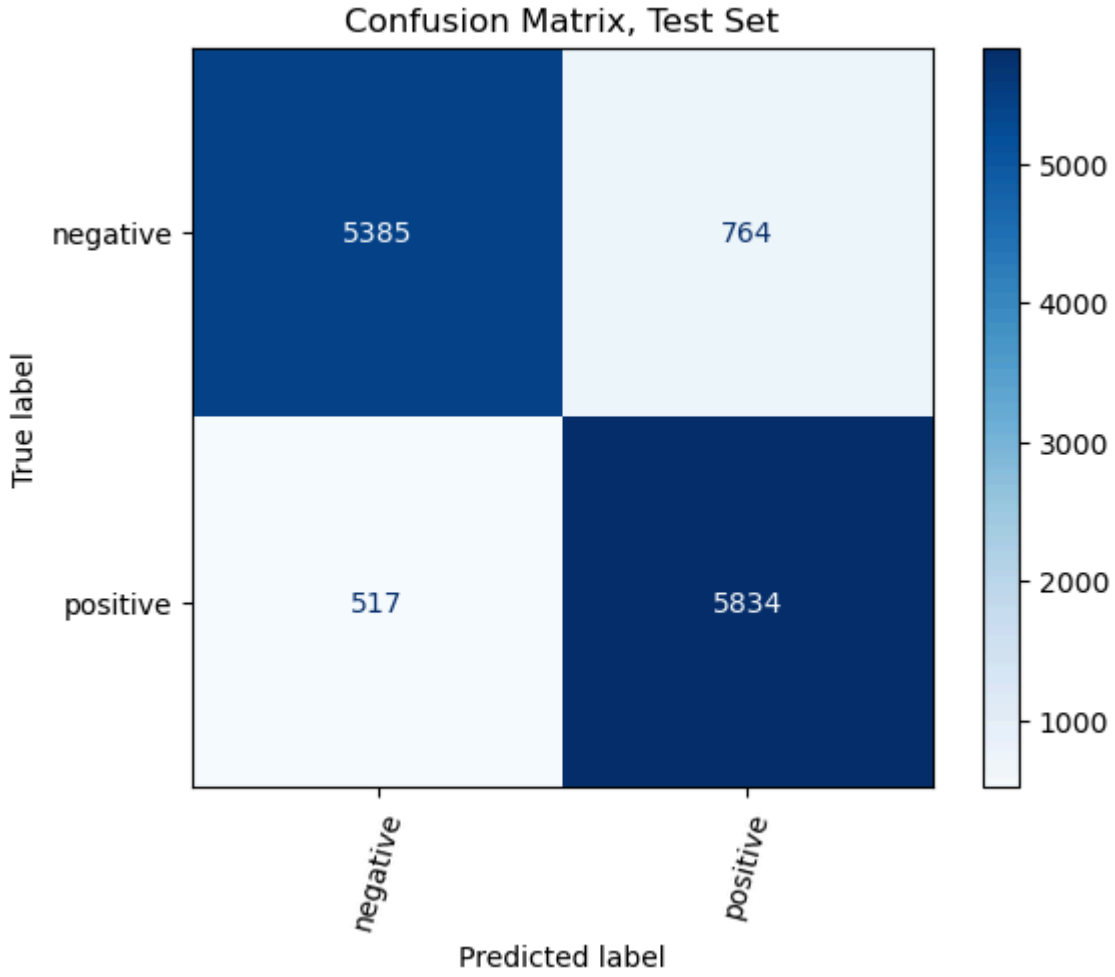


SVM:

- **F1 Score:** 0.8975
- **Confusion Report:**

	precision	recall	f1-score	support
negative	0.9124	0.8758	0.8937	6149
positive	0.8842	0.9186	0.9011	6351
accuracy			0.8975	12500
macro avg	0.8983	0.8972	0.8974	12500
weighted avg	0.8981	0.8975	0.8974	12500

- **Confusion Matrix:**



ERROR ANALYSIS

As we can see, even though we have good scores, some even **nearing 0.90**, we still see a small fraction of texts that were misclassified, to get a better grip of why, we will perform a small analysis on those samples:

In some cases, reviews included **both positive and negative** comments, suggesting that a **neutral** category might be useful for this type of review. Other misclassified reviews contained words like **“weird,” “not,” “impatient,”** or **“evil,”** which, if picked up by the CountVectorizer, could lead the models to incorrectly classify them as **negative**. Or, in the contrary, the **negative** reviews contain words like **“good”, “success”** or even words like **“interesting” and "unbelievingly",** however, what the model didn´t pick up was the presence of the word **“not”** just before them, classifying them as **positive**, when in reality, it’s quite the contrary.

index	original_text	true_label	predicted_label
21768	Why a good actress like Elizabeth Berkley stars in this commonplace movie???!!! The cast gives some good performance (Elizabeth Berkley as a Barbie girl, Ele Keats as a girl without mother and Justin Whalin, a guy eternally lessened by his bother), but the direction is extremely boring and the story is NOT so interesting and original. I can NOT believe that a movie like this was produced for the big screen! Julie Corman (the producer): are you CRAZY???!!!	positive	negative
42560	God cuts himself with a straight razor, afterwards, he gives birth to mother earth, which then gives birth through gods semen, son of earth. This is a one and a half-an-hour movie that sort of depicts a dark version of how god created the world. Its surreal, dark and poetic. The most important aspect of the film to me is the visuals. Its shot in a very grainy black and white film, using both slow-motion and normal shooting. Sometimes even fast film and stop-motion. The scenes are long and dragged out, that sets a very weird mood. No sounds were recorded when filming (i think), sound effects were added afterwards, such as criccets, water and other ambient sounds, repeated over and over. As you might understand this movie is certainly not for the impatient person... I often felt it was similar to David Lynch's Eraserhead, only this one is even harder to understand, and even more dragged out, but that's okay for me, I like that kind of stuff. Its certainly not entertainment (at least not the Hollywood kind) so if your going to check it out, i recommend you set yourself up for it. Be open minded, and except something like	positive	negative
44149	This documentary focuses on the happenings in Gothenburg 2001. In swedish media the demonstrators where pictured as criminals that stood for anarchy and violence. This movie shows that there not, actually they are intelligent, articulate people that has something to say - And says it by the force of bricks. They believe in a better world, a world where people can think and say what they want - without being aimed at. But are their beliefs of having the possibility of changing the society realistic? I think not. This documentary gives us enlightenment in the issues of greed, capitalism and the future it might bring. It is a great documentary that is not propaganda, because it is not shown as what they say is right. Everything it shows is what some individuals think and it is up to the viewer to decide if what they do and stand for is right or wrong. I have heard many people that labels this a propaganda and therefor chooses not to view it, I think they are making a bad decision because even if you sympathize with the police or the demonstrators belief, all you get is more facts to rely on, for example the kid that got shoot says that he thinks that it is good to throw a brick through a McDonalds window because it is the step between thinking and acting as you think. Overall this movie freaked me out because you cannot really dismiss the facts that the few policemen, that fought violence with violence, did not get convicted or even detained in custody even however the proof of them throwing bricks at the demonstration march (and in some cases beating people with truncheon, even though they are lying on the street without making resistance) where as good as it gets. Rating: 8/10 - Very good, but not best!	positive	negative
42538	Though I had sort of enjoyed THE SATANIC RITES OF Dracula (1974), I knew I shouldn't expect too much from its even more maligned predecessor! Surely the least of the Hammer Draculas (Marcus Hearn on the Audio Commentary for THE CREEPING FLESH [1973] even goes so far as to call it the studio's nadir!), the film really flounders due to its totally unhip - and now embarrassingly dated - updating of the myth (the modern-day setting actually suited SATANIC RITES rather better)...even if, truth be told, it's still vastly preferable to dreck like Dracula 2000 (2000) or VAN HELSING (2004)! Despite Christopher Lee's vociferous bashing of the film, he still cuts an undeniably striking figure as the undead vampire (even if he appears very little and is inexplicably confined to one setting); likewise, Peter Cushing delivers his usual committed performance. The only other noteworthy acting job in the film is that given by Christopher Neame (son of director Ronald) as Johnny Alucard(!) - even if that's only because of how unbelievably hammy it is! Unfortunately, the two best-known female members of the cast (both of them horror regulars) - Stephanie Beacham and Caroline Munro - can't rise above their physical attributes. The camera-work is by Dick Bush (who had shot THE BLOOD ON SATAN'S CLAW [1971] and, for Hammer, WHEN DINOSAURS RULED THE EARTH [1970] and TWINS OF EVIL [1971] - but is perhaps best known for his longtime association with Ken Russell) which manages some nice atmosphere throughout, especially during three crucial sequences: the carriage-ride scuffle at the (properly Gothic) beginning; the hysterical Black Mass sequence, followed by the resuscitation of Dracula; and the final confrontation between Lee and Cushing's Van Helsing.	negative	positive

These two were extracted from the Logistic Regression.

index	original_text	true_label	predicted_label
21768	Why a good actress like Elizabeth Berkley stars in this commonplace movie???!!! The cast gives some good performance (Elizabeth Berkley as a Barbie girl, Ele Keats as a girl without mother and Justin Whalin, a guy eternally lessened by his bother), but the direction is extremely boring and the story is NOT so interesting and original. I can NOT believe that a movie like this was produced for the big screen! Julie Corman (the producer): are you CRAZY???!!!	positive	negative
46174	i got to see the whole movie last night and i found it very exciting.it was at least,not like the teen-slasher movies that pop out every now and then.the search for the killer and the 'partner' relationship between the hero&the so-called bad guy was parts i liked about the movie.also,i remember once being on the edge of my seat during a specific scene in the movie.i mean it's exciting.maybe some time later,i might watch the movie again...	positive	negative
31080	You have to hand it to writer-director John Hughes. With enormous success behind him in the misfit-teenager/high school vein, he managed to branch out into other areas of comedy, finding in the bargain a great ally in comedian-actor John Candy. Here, goof-off adult Candy becomes a better person after agreeing to babysit his brother's wiseacre kids; it's a surefire formula designed to please both cynical teens as well as their parents, and it isn't any wonder the film was a winner with theater audiences. Still, Hughes relies almost completely on Candy's charm to put the scenario over, and one may eventually grow tired of the repetitious gags with the star front and center. The kids are sitcom-smart, the other adults shapeless blobs, and Amy Madigan is too intense, too hyped-up playing Uncle Buck's girlfriend. Later became a TV series, which is befitting since the material was already television-perfected. *1/2 from ****	negative	positive
17949	I just saw Adam Had Four Sons for the first time and the thing that struck me was that I believe that the model used was Theodore Roosevelt and his four sons. They were approximately the same ages as the four boys in this film. Warner Baxter in his portrayal of Adam Stoddard talked about the same values and family tradition that you would have heard from our 26th president without some of the more boisterous aspects of TR's character. Like TR all of the Stoddard sons serve in World War I, in this case though the youngest only loses an eye instead of being killed. But what if a female minx gets into this all male household and disrupts things? That's Susan Hayward's job here. In one of her earliest prominent roles, Hayward is a flirtatious amoral girl who marries one son, has an affair with another, and starts making a play for the third. It's an early forerunner of the kind of a part that later brought her an Oscar in I Want to Live. I suppose that with as powerful a model of decorum as Theodore Roosevelt was and Warner Baxter portrays, everyone is afraid to tell Father what's going on. The sons and also their governess Ingrid Bergman. Here's where the plot gets a little silly. Bergman is introduced to us as a governess hired by Baxter and wife Fay Wray for their kids. Wray dies and Baxter suffers some financial reversals in business. Bergman has to be let go. She goes back to France and years later comes back to the family when the kids are grown up. I'm sorry, but I can't believe the kids need a governess now. Hayward is quite right when she confronts her that it wasn't the kids who brought her back. In the normal course of things, Bergman would have gotten on with her life. One of the previous reviewers said that a quarter to a third of the film I have was edited out. Possibly that could be the reason for the many plot holes we have. It's too bad that Ingrid and Susan could not have done another film together in the Fifties when Hayward was at her heights and Bergman had just made a comeback. Susan Hayward is the main reason to see Adam Had Four Sons. And I'm willing to believe that a good deal of Ingrid was left on the cutting room floor.	negative	positive
43224	I agree with one of the other comment writers about good story & good actors but mismatched, and I would also say rushed. It has been	negative	positive

These were extracted from the NB.

	original_text	true_label	predicted_label
24006	this movie is extremely funny and enjoyable,with suitable, funny and experienced casts. I find this movie enjoyable not only by the elements of humor but also the music in various scenes. Kevin Kline, a good comedian has done a good job at being funny in many parts of the film along with Tom Selleck who is amazingly different from many of his other films. The humor within this film are goofy which makes various exaggerations within many scenes, especially the beginning bits. Joan Cusack is also remarkably funny and exaggerated; and the same goes for all the other casts. This film has many elements of goofy humor and is enjoyable if you want to laugh.	positive	negative
3935	Garam Masala is one of the funniest film I've seen in ages. Akshay Kumar is excellent as the womaniser who has affairs with 3 girls and engaged at the same time. John Abraham is Amusing at times and this is one of his best works so far. Paresh Rawail is superb as usual in most of his films. The director Priyadarshan has delivered great Movies in the past. Hera Pheri, Hungama and Hulchul being some of the Best. Garam Masala is his funniest film he has made. The three newcomer actresses are average. Rimi sen doesn't get much scope in this movie. I was impressed to see how Priyadarshan made a movie with a simple storyline of a guy having a affair with 3 girls at the same time. All 3 girls have a day off in the same day and end up in the same house. Packed with loads of Laughs, this is one Non stop Entertainer.	positive	negative
21768	Why a good actress like Elizabeth Berkley stars in this commonplace movie???!!! The cast gives some good performance (Elizabeth Berkley as a Barbie girl, Ele Keats as a girl without mother and Justin Whalin, a guy eternally lessened by his bother), but the direction is extremely boring and the story is NOT so interesting and original. I can NOT believe that a movie like this was produced for the big screen! Julie Corman (the producer): are you CRAZY???!!!	positive	negative
46174	i got to see the whole movie last night and i found it very exciting.it was at least,not like the teen-slasher movies that pop out every now and then.the search for the killer and the 'partner' relationship between the hero&the so-called bad guy was parts i liked about the movie.also,i remember once being on the edge of my seat during a specific scene in the movie.i mean it's exciting.maybe some time later,i might watch the movie again...	positive	negative
28958	Loved this film. Real people, great acting, humour, unpredictable. The characters were believable and you really connected with them. If you're looking for a film about slightly offbeat characters outside the mainstream of society and how they help each other, this would be a good choice.	positive	negative

These were extracted from the SVC.