Understanding IST736 Text Mining - HW4 - Multinomial Naïve Bayes

Use Python [Weka is optional]
Multinomial Naïve Bayes algorithm
Data:
(1) sentiment (positive or negative)
(2) authenticity (true or fake/false, lie detection)

OK – let's start by thinking about what this is all about. I am going to walk you through what you should do each time you read an assignment or similar work project that uses data.

First, let's visually (with our eyes) look at the dataset itself – what are we dealing with here?

Well, the first thing we see is that we have "labeled data". In fact – we have TWO labels. What are they? How can you tell?

If you said "lie" and "sentiment" you are right!

What do they mean?
Well, "lie" has two levels or categories: "f" and "t".

If we are not told on the data sheet what f and t are – we have to use our common sense. The letter f, especially when used with the letter t often stand for false and true (or something similar).

The column name is called "lie". So, "f" means the review is false or fake (a lie) and t means the review is true (not a lie). OK – so we have true and false. (False and Fake are the same)

Next, we have a label called sentiment with "n" and "p". Used together, these generally mean n for negative and p for positive. So we will assume that a review with an "n" is a negative review and a review with a "p" is a positive review.

This is a good start!

What are these reviews of?
Let's look at a few of the words in a few random reviews:

> **'Mike\'s Pizza High Point**
> **'i really like this buffet restaurant in Marshall street. they have a lot of selection of american**
> **'OMG. This restaurant is horrible. The receptionist did not greet us**

OK – that did not take long. These are restaurant reviews!

So, some are real (t for true) and some are fake (f for false/fake) and some are positive (p) and some are negative (n).

What is the TOPIC of this?

1) **Is it naïve bayes?** No – that's the model we will use.

2) **Is it sentiment analysis?** – Not really – that is the method – and we will discuss that in the Analysis section.
3) **OK – so what is the TOPIC??** It's restaurant reviews! Right? This is about restaurant reviews. It is not about global warming or SAT test scores. The topic here is restaurant reviews.

What will the Introduction be like?

The Introduction will start out talking about how Americans love to eat and love to eat out at restaurants. A fun fact might be how much MONEY is spent on this each year and how often an average person eats out. This can then lead into how people choose restaurants and the idea that restaurant reviews are becoming more and more critical – not only for choices that people make, but for restaurant owners…. (see where I am going with this….)

Next, talk about how restaurant reviews work (not technically – but rather so a 9 year old can clearly relate). From here, lead into how the nature of reviews (fake or real and pos or neg) can really affect people's choices and restaurant success. Is it possible that people choose a restaurant based on positive reviews and reject a restaurant based on negative reviews?

Does this make sense?

Can you see how the above is introducing and motivating the topic here?

You can also include some fun images – perhaps one of a restaurant review ☺

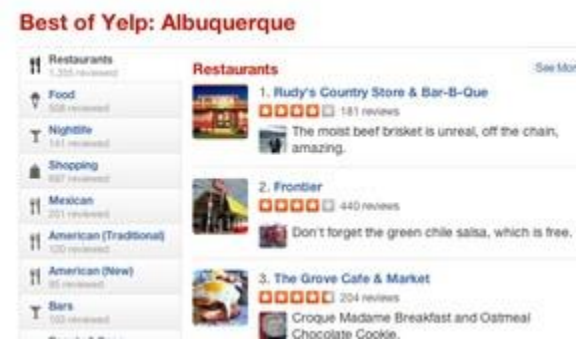

Figure X: Restaurant reviews…

In the Introduction, you can also talk about the dangers of fake reviews (lies). Who is affected by this and why?

Finally, lead into the idea that reviews can be evaluated for honesty and sentiment (public opinion) and that this information can be useful for restaurant-goers and restaurant owners…

That's an example of what goes into an introduction.

The conclusions will review (but also answer and address possible solutions) for some or all of the above based on the analysis and results.

----------NEXT----------

What goes into the Analysis?

OK – so where are we? We have a great Intro (which should be MORE than what I noted above). Notice the Introduction does not discuss the dataset or the data cleaning, or the fact this is your assignment ☺

Now it's time to look at the data again – but this time – using Python.

We can see that the data has decent and clear labeling, but the csv is ugly and needs to be cleaned.

The **first subsection of Analysis** (which should be many pages (not paragraphs) in length) will be ***Data Cleaning, Prep, Formatting, and EDA***.

A BEFORE and AFTER Figure is great here.

DO NOT copy and paste Python output into your paper. This is a no-no in real life. Instead, interpret and create.

So, here is a nice image of my BEFORE data:

The Raw Dataset: A quick look at the format

lie	sentiment	review			
f	n	'Mike\'s P	NY Servic	not. Stick to pre-made dishes like stuffed	
f	n	'i really lik	japanese	and chinese dishes. we also got a free dri	
f	n	'After I we	we went to DODO restaurant for dinner. I found wo		
f	n	'Olive Oil	and the waitor had no manners whatsoever. Don\'t		
f	n	'The Sever	never more. '		
f	n	'I went to	she rudel	I had a ba	otherwis	my favorite place to
f	n	'I went to	our food	especially if you\'re in a hurry!'	
f	n	'I went to	and then	at which point we just paid and left. Don\	

Figure 1: The raw data – a small example taken from the larger file.

As you use Python to clean the data, note (discussion, explain, and describe) any key steps in the Data Cleaning subsection, use vis as needed, and at the end, **create a similar figure of the clean data – the AFTER**.

Discuss stopwords - what did you use?
How did you decide on them?
How did you deal with punctuation?
Cases?
Tabs? (these are the \t) – etc – etc.
Be transparent and clear.
Show examples when they are needed.
Use vis.
Discussion things that were interesting, unexpected, etc.

OK! Now that the data is cleaned and the first subsection of the Analysis is created – its time to move on to the ***Models and Methods subsection of Analysis…***

What are the models and methods?

In this case, you are asked to:

For each of the two classification tasks, use MNB to build the models, and evaluate them using 10-fold cross validation methods.

This is actually a lot of stuff….let's break it down:

1) There are TWO classifications here – one for LIE and one for SENTIMENT. You will need TWO datasets! You will need to build this.
2) You will be using Naive Bayes in Python (yes Weka is optional – I recommend Python).
   Naïve Bayes is your MODEL.
3) Naïve Bayes is a supervised learning method that required labeled data (which we have) and can be tested for accuracy using confusion matrices and cross validation….

Let's break this down further…

The Analysis part - subsection Methods and Models – will start with an explanation of what Naïve Bayes is.
How does it work?
What are the requirements and assumptions for Naïve Bayes?
Are there data requirements – such as labeled, independence, qualitative, quantitative, etc. etc.

Once you discuss NB, next build the model.

You will need **testing and training data.**
You will be using **10-fold cross validation**.

Discuss both!
Not sure what cross validation is? (View the class lectures and then look it up online_.
Find a great visual illustration of cross-validation and include it in your paper.

Show a figure (vis) of a portion of the testing set and show a figure (vis) of a portion of the training set.

Have an image or vis that illustrates what cross validation is.
In other words, part of the Analysis section is to explain, describe, and illustrate the methods and models.

Should you show actual code? No – never unless you are teaching the reader how to code (which in this class you are not)

Can you have a figure of a small portion of the testing and training sets (post cleaning and prep)? Yes!

Should you have several testing and training sets in this case?

Well – let's think about that….

How many labels do you have? (two!)

So, you will need at least one testing and training set for Label 1 (lie) and one testing and training set for label 2 (sentiment). Will the data be the same? Sure! It's the label that will be different.

Can you keep both labels in the datasets used to build the models? Nope! Why not?

How about normalization – should you also have other testing and training sets that are normalized? Sure!

In fact, the Analysis section and all the options can and should be many pages long and show a true engagement in the methods, models, options, and creation, etc.

So, where are we now? At this point, the Introduction is written, the data is clean and ready to use, and all of the above is complete.

Now we are ready for the Results.

But first – its time to do some more coding. So far, we have clean data and we have several variations of testing and training sets for each label and some that are normalized and some not.

We now need to code Naïve Bayes and test out various datasets.

Each dataset will have its own subsection in Results.

What?? What do you mean "each dataset"??

Well, remember, the data we are given is NOT one dataset. It is two (at least). One for LIE and one for SENTIMENT. Next, if you normalize, that will create other datasets.

In the Results section, you will also discuss and illustrate the results.
How?
Using confusion matrices, discussions, comparison tables, and vis.

So, let's clarify this. Let's really break this down. What do we really mean by "the different data sets??"

First, to answer that, let's look again at the raw data. Here is a small figure:

lie	sentiment	review						
f	n	'Mike\'s P	NY Servic	not. Stick to pre-made dishes like stuffed				
f	n	'i really lik	japanese	and chinese dishes. we also got a free drir				
f	n	'After I we	we went to DODO restaurant for dinner. I found wor					
f	n	'Olive Oil	and the waitor had no manners whatsoever. Don\'t g					
f	n	'The Seve	never more. '					
f	n	'I went to	she rudel	I had a ba	otherwis	my favorite place to c		
f	n	'I went to	our food	especially if you\'re in a hurry!'				
f	n	'I went to	and then	at which point we just paid and left. Don\'				

This data actually contains two base data sets:

	A	B
	lie	review
	f	'Mike\'s Pizza High Point
	f	'i really like this buffet restaurant in Mars
	f	'After I went shopping with some of my fr
	f	'Olive Oil Garden was very disappointing.
	f	'The Seven Heaven restaurant was never
	f	'I went to XYZ restaurant and had a terribl
	f	'I went to ABC restaurant two days ago an
	f	'I went to the Chilis on Erie Blvd and had t
	f	'OMG. This restaurant is horrible. The rece
	f	Yesterday
	f	'Last weekend I went to a place called Rat

sentimen	review		
n	'Mike\'s P	NY Servic	not. Stick to pre-made dishes like st
n	'i really lik	japanese	and chinese dishes. we also got a fr
n	'After I we	we went to DODO restaurant for dinner. I foun	
n	'Olive Oil	and the waitor had no manners whatsoever. D	
n	'The Seve	never more. '	
n	'I went to	she rudel	I had a bad experience as I had take
n	'I went to	our food	especially if you\'re in a hurry!'
n	'I went to	and then	at which point we just paid and left.
n	'OMG. Thi	we just st	the crust of a pizza was so hard like
n	Yesterday	I went to	I thought its just a name but later at
n	'Last week	and I was	but it seems like everyone around h
n	'I once we	I found th	there is no sauce! No any kind of sa
n	'I entered	said coldl	help yourself. Then she just disappe

Once these two datasets are cleaned, tokenized, vectorized, etc. They can each be "normalized". This will produce two more datasets. The normalized LIE dataset and the normalized SENTIMENT dataset.

Also, there is more than one way to normalize. If you choose a second normalization method, that will create more datasets, and so on.

Common student question: How many ways to we "have to" normalize?? As many ways as you wish. You are the professional wishing to become a data scientist.

Kind suggestion: Stop thinking about how little can be done and think about how much can be done.

Each dataset you have must now be split into **test and train data** – why?

Once you have testing and training data (for each dataset you plan to investigate), you can now train and then test using Naïve Bayes.

The Results section will:

- show
- discuss
- visualize and
- offer a confusion matrix (for each trained model)
- and calculate accuracy for all the different datasets.

A good way to organize this is to create subsections:

Results

Model 1: Lie Data (not normalized)

Model 2: Lie Data (normalized with TF-IDF or other)

Model 3: Sentiment Data (not normalized)

Model 4: Sentiment Data (normalized with…)

For each model, you will include a description and show figures of small portions of the testing and training data (so the reader can easily see).

For each model and dataset, you will have a wordcloud, you will run Naïve Bayes, you will have and discuss the confusion matrix, you will calculate and discuss the accuracy, etc.

You will also:

Report the 20 most indicative words that the models have learned.

What is a good way to SHOW this? How about get the frequencies of all the words. Sort from highest to lowest and create a table and a word cloud to show these.

Based on these words, do you think the models have learned the concepts (lie or sentiment) that they are expected to learn?

In other words – how well did your models do?
Where they accurate?
What does the confusion matrix show?
What does the confusion matrix suggest?

What happens if you remove more words (your choice which ones based on your observations and results)? Etc.

As you can see, the Results section is many pages long and filled with detailed investigation, comparison, and results illustration and discussion. In the "real-word", the Results part would go on (iteratively) for a long time. You would try – review – evaluate – fix – try – review – evaluate – fix – until you discovered what was really going on and how to improve it. The more of this you can do – the better.

Finally, its time to understand what happened.

What did happen??

- Were you able to model and then predict a lie?
- Were you able to model and predict sentiment?
- Was this what you expected?
- Did normalization affect the results?
- What one was easier to model?
- Why do you think this is?

All of the above can and should go into Results – as this is all still technical!

Now – finally – we get to write the Conclusions. Your Results will have told you whether is it is feasible to predict and model lies. But, in Conclusions, you do not want to repeat the Results or be technical. Instead, you want to bring this down to a human level.

Can restaurant reviews be trusted. Should we, as restaurant goers, use online reviews to make decisions about which restaurant to choose?

Can online reviews help restaurant owners? Should restaurant owners be aware of "fake" reviews?

Etc….

Can you see how the Conclusions are all about the TOPIC and how it affects real-life.

The above was a complete and step by step overview of how to go about competing an assignment using data.

A few reminders:

1) Never use "I", "we", "you", etc.
2) Create the template and outline first – and then fill it in. This will help to organize you and direct you.
3) Work smart! For Analysis and Results – do this all first for a small sample (using only sentiment or only lie) – just be make sure the code works. You can even build clean data first, get Naïve Bayes to work etc, and then clean the real data.

TIME: There is no way to do this (or learn this) without spending the time. There is no magic trick – no short-cut – etc.