Better News through Machine Learning

Casey Stella

September 21, 2011

Table of Contents

Statement of Problem

- News in the internet-age is decentralized

Statement of Problem

- News in the internet-age is decentralized
- This is good

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias
  - "We report, you decide"

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias
  - "We report, you decide"
- I want to automatically determine if text has a political slant.

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias
  - "We report, you decide"
- I want to automatically determine if text has a political slant.
  - This is a very broad problem.

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias
  - "We report, you decide"
- I want to automatically determine if text has a political slant.
  - This is a very broad problem.
  - This is a very **hard** problem.

Statement of Problem

- News in the internet-age is decentralized
- This is good
  - More voices means more perspectives
  - Greater access means more more refined coverage
- This is also bad
  - It's hard to detect bias
  - "We report, you decide"
- I want to automatically determine if text has a political slant.
  - This is a very broad problem.
  - This is a very **hard** problem.
  - This is a very **vague** problem.

Challenges

- Need to extract text from HTML

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java
- Need to classify text as political or apolitical

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java
- Need to classify text as political or apolitical
- Need to classify political text as left-leaning, right-leaning or centrist

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java
- Need to classify text as political or apolitical
- Need to classify political text as left-leaning, right-leaning or centrist
  - These categories are vague and inherrently subjective

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java
- Need to classify text as political or apolitical
- Need to classify political text as left-leaning, right-leaning or centrist
  - These categories are vague and inherrently subjective
  - Need to make them the least subjective as possible

Challenges

- Need to extract text from HTML
  - This is taken care of for us by the Boilerpipe library in Java
- Need to classify text as political or apolitical
- Need to classify political text as left-leaning, right-leaning or centrist
  - These categories are vague and inherrently subjective
  - Need to make them the least subjective as possible
- Be as lazy as possible

A Model for Political Orientation

- This can be tackled with NLP classification techniques

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias
- "Bias" is difficult to characterize

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias
- "Bias" is difficult to characterize
  - One approach is to map politicians onto a 1-D spectrum and segment the specrum into left, right and center

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias
- “Bias” is difficult to characterize
  - One approach is to map politicians onto a 1-D spectrum and segment the specrum into left, right and center
  - Use the speeches from the politicians as samples

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias
- "Bias" is difficult to characterize
  - One approach is to map politicians onto a 1-D spectrum and segment the specrum into left, right and center
  - Use the speeches from the politicians as samples
  - All that is left is determining relative position on the 1-D spectrum and gathering the data

A Model for Political Orientation

- This can be tackled with NLP classification techniques
- Need a sample of politically oriented text segmented by political bias
- "Bias" is difficult to characterize
  - One approach is to map politicians onto a 1-D spectrum and segment the specrum into left, right and center
  - Use the speeches from the politicians as samples
  - All that is left is determining relative position on the 1-D spectrum and gathering the data
- Thankfully, I found a dataset with speeches from senators from the 111^{th} Congress

Computational Political Science

- Computational Political Science to the Rescue

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.
- Senators and bills are fitted to the 1-D spectrum using logistic regression

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.
- Senators and bills are fitted to the 1-D spectrum using logistic regression
  - The fitting is such that a senator's proximity to a bill is proportional to their probability for voting 'Yay' on the bill

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.
- Senators and bills are fitted to the 1-D spectrum using logistic regression
  - The fitting is such that a senator's proximity to a bill is proportional to their probability for voting 'Yay' on the bill
  - This provides an ordering that groups senators by voting record

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.
- Senators and bills are fitted to the 1-D spectrum using logistic regression
  - The fitting is such that a senator's proximity to a bill is proportional to their probability for voting 'Yay' on the bill
  - This provides an ordering that groups senators by voting record
- The hard statistics is done for me by the good people at voteview.com

Computational Political Science

- Computational Political Science to the Rescue
- The idea is to use roll-call votes to fit senators onto the 1-D spectrum from left-to-right.
- Senators and bills are fitted to the 1-D spectrum using logistic regression
  - The fitting is such that a senator's proximity to a bill is proportional to their probability for voting 'Yay' on the bill
  - This provides an ordering that groups senators by voting record
- The hard statistics is done for me by the good people at voteview.com
- Obviously the model is simplification, but for the purpose of this project, we'll pretend it's a pretty good model.

Machine Learning

- Now we have a set of documents associated with political orientations

Machine Learning

- Now we have a set of documents associated with political orientations
- We can split the dataset into a training set and testing set and evaluate different machine learning algorithms

Machine Learning

- Now we have a set of documents associated with political orientations
- We can split the dataset into a training set and testing set and evaluate different machine learning algorithms
- Tried many algorithms, but the ones that worked best was Adaptively Boosted Decision Trees

Machine Learning

- Now we have a set of documents associated with political orientations
- We can split the dataset into a training set and testing set and evaluate different machine learning algorithms
- Tried many algorithms, but the ones that worked best was Adaptively Boosted Decision Trees
- Decision Tree classifiers "learns" a decision tree by being presented with many examples from a set of categories. The leaves of the trees are categories and the interior nodes are input variables.

Machine Learning

- Now we have a set of documents associated with political orientations
- We can split the dataset into a training set and testing set and evaluate different machine learning algorithms
- Tried many algorithms, but the ones that worked best was Adaptively Boosted Decision Trees
- Decision Tree classifiers "learns" a decision tree by being presented with many examples from a set of categories. The leaves of the trees are categories and the interior nodes are input variables.
- This is a weak classifier, but can be boosted by creating a meta-learning algorithm on top called adaptive boosting

Evaluation of Bias Classifier

- I chose the middle $\frac{5}{8}^{th}$ of the data to be my center

Evaluation of Bias Classifier

- I chose the middle ${\frac{5}{8}}^{th}$ of the data to be my center
- Total Accuracy (95% confidence) is $78\% \pm 0.04$

Evaluation of Bias Classifier

- I chose the middle $\frac{5}{8}^{th}$ of the data to be my center
- Total Accuracy (95% confidence) is $78\% \pm 0.04$
-

		Predicted			
		Left	Center	Right	Total
Actual	Left	46(69%)	16(24%)	4(6%)	66
	Center	27(10%)	202(78%)	29(11%)	258
	Right	0(0%)	7(11%)	52(88%)	59

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.
- Topic Models can be generated from a corpus of documents

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.
- Topic Models can be generated from a corpus of documents
  - The best known topic model is Latent Dirichlet Allocation

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.
- Topic Models can be generated from a corpus of documents
  - The best known topic model is Latent Dirichlet Allocation
  - Topic models create a set of vectors representing the topics in the corpus

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.
- Topic Models can be generated from a corpus of documents
  - The best known topic model is Latent Dirichlet Allocation
  - Topic models create a set of vectors representing the topics in the corpus
  - New documents can be represented as linear combinations of topics where the coefficients represent the degree to which a topic contributes to the document

Classifying Text as Political/Apolitical

- We only want to look for bias in political texts, so we need to know which texts have political content.
- Topic Models can be generated from a corpus of documents
  - The best known topic model is Latent Dirichlet Allocation
  - Topic models create a set of vectors representing the topics in the corpus
  - New documents can be represented as linear combinations of topics where the coefficients represent the degree to which a topic contributes to the document
  - Such as, consider topics v_1 and v_2 which represent roughly "healthcare" and "the war in iraq", you can represent a story about hospitals in the warzone as $0.2v_1 + 0.9v_2$ and a story about a hospital closing as $0.8v_1 + 0v_2$

Classifying Text as Political/Apolitical

- We can generate a topic model from the corpus of senatorial speeches

Classifying Text as Political/Apolitical

- We can generate a topic model from the corpus of senatorial speeches
- This gives us a vector space and a way to map documents onto it

Classifying Text as Political/Apolitical

- We can generate a topic model from the corpus of senatorial speeches
- This gives us a vector space and a way to map documents onto it
- Now we can use distance metrics to construct an inclusion/exclusion criteria for political documents

Classifying Text as Political/Apolitical

- We can generate a topic model from the corpus of senatorial speeches
- This gives us a vector space and a way to map documents onto it
- Now we can use distance metrics to construct an inclusion/exclusion criteria for political documents
- Roughly, define a metric $||\cdot||$ and a real number k such that $||\vec{v}|| < k$ implies that the document is political for any document $\vec{v}$.

Classifying Text as Political/Apolitical

- We can generate a topic model from the corpus of senatorial speeches
- This gives us a vector space and a way to map documents onto it
- Now we can use distance metrics to construct an inclusion/exclusion criteria for political documents
- Roughly, define a metric $||\cdot||$ and a real number k such that $||\vec{v}|| < k$ implies that the document is political for any document $\vec{v}$.
- The trick now becomes defining $||\cdot||$.

Mahalanobis Distance

- There are statistical distance metrics which give us the rough distance from a given dataset's "center of mass"

Mahalanobis Distance

- There are statistical distance metrics which give us the rough distance from a given dataset's "center of mass"
- **Mahalanobis distance** is just such a distance metric

Mahalanobis Distance

- There are statistical distance metrics which give us the rough distance from a given dataset's "center of mass"
- **Mahalanobis distance** is just such a distance metric
- We have a set of documents and their respective vectors, so we can define a distance function to be the distance from this set

Mahalanobis Distance

- There are statistical distance metrics which give us the rough distance from a given dataset's "center of mass"
- **Mahalanobis distance** is just such a distance metric
- We have a set of documents and their respective vectors, so we can define a distance function to be the distance from this set
- So all documents who have vectors with a sufficiently large Mahalanobis distance contain topics that are dissimilar to the corpus of political speeches.

Mahalanobis Distance

- There are statistical distance metrics which give us the rough distance from a given dataset's "center of mass"
- **Mahalanobis distance** is just such a distance metric
- We have a set of documents and their respective vectors, so we can define a distance function to be the distance from this set
- So all documents who have vectors with a sufficiently large Mahalanobis distance contain topics that are dissimilar to the corpus of political speeches.
- Unfortunately, I haven't gotten around to evaluating this approach.

What did we learn?

- Using good libraries makes hard problems much easier

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct
  - I think this may be due to training on clean data and evaluating on noisy data

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct
  - I think this may be due to training on clean data and evaluating on noisy data
  - Also, arbitrary text from the internet isn't the same style as political speeches from senators

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct
  - I think this may be due to training on clean data and evaluating on noisy data
  - Also, arbitrary text from the internet isn't the same style as political speeches from senators
- Machine Learning is like a wolverine on a leash.

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct
  - I think this may be due to training on clean data and evaluating on noisy data
  - Also, arbitrary text from the internet isn't the same style as political speeches from senators
- Machine Learning is like a wolverine on a leash.
  - Once you let it go, you're never quite sure what it's going to do or when it's going to turn on you and eat your face.

What did we learn?

- Using good libraries makes hard problems much easier
- I think I might have solved the wrong problem
  - When evaluating real data, my classifier sometimes doesn't match my gut instinct
  - I think this may be due to training on clean data and evaluating on noisy data
  - Also, arbitrary text from the internet isn't the same style as political speeches from senators
- Machine Learning is like a wolverine on a leash.
  - Once you let it go, you're never quite sure what it's going to do or when it's going to turn on you and eat your face.
- Cleaning data is important

Questions

Thanks for your attention! Questions?

- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cestella@gmail.com

Questions

Thanks for your attention! Questions?

- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cestella@gmail.com
- Oh, and by the way, Explorys is hiring!