Accelerated Natural Language Processing
Lecture 10
Dialect and discrimination

Sharon Goldwater

7 October 2019

School of informatics

A change of pace

So far, focused on technical content:

- Linguistics: properties of language, ambiguity, etc
- Computation: models and tasks (LM, tagging, classification)

Today, zoom out to "bigger picture":

- Where does our data come from and why does that matter?
- What effects might our research/products have on society?

Further lectures on related topics later in the course.

Today's lecture

- What is algorithmic bias and what are some examples?
- What are some ethical and legal implications?
- What is a dialect and why is dialect variation increasingly important in NLP?
- How can dialect variation lead to racial discrimination in NLP systems?

The role of subjectivity

Easy to think science and engineering are objective.

- We parse 3 sentences per second.
- Our system gets 97% accuracy on the test set.

The role of subjectivity

Easy to think science and engineering are objective.

- We parse 3 sentences per second.
- Our system gets 97% accuracy on the test set.

But lots of aspects are subjective.

- Is speed more important than accuracy?
- Which research questions are interesting or important?
- How should we design the study or interpret the results? (Whose data or viewpoints are included?)

The answers often have **ethical implications**.

Ethical implications?

That is, moral issues of right and wrong, often involving both potential **benefits** and **risks or disadvantages**. For example,

- Studying one disease could mean less funding for another.
- Data collection might improve user experience, but reduce privacy.
- Making a system more usable for most people might make it less usable for others.

Ethical implications?

That is, moral issues of right and wrong, often involving both potential **benefits** and **risks or disadvantages**. For example,

- Studying one disease could mean less funding for another.
- Data collection might improve user experience, but reduce privacy.
- Making a system more usable for most people might make it less usable for others.

The "right" answer is highly personal, but shaped by culture and experience.

- This lecture shaped by my own experience. Yours may differ.

Why is this discussion important?

- Fields such as medicine and psychology have long-standing ethical guidelines and education.
- Computer science does not have this history, and until recently AI had little impact on broader society.
- Result: researchers and designers often haven't considered potential ethical problems until it's too late.
- My goal is to raise your awareness so you'll consider risks **early**, and hopefully mitigate them.
  - Not just me: AI Fairness, Accountability, Transparency (FAT) is now a big deal! Workshops on FAT ML, Ethics in NLP, etc.

Focus today: algorithmic bias

When an algorithm's outputs differ systematically and unfairly between one group of people and another.

- Famous example: face recognition systems.[1]
  - Have a harder time detecting Black faces than White faces, and are more likely to falsely detect a match between Black faces.
  - Annoying if you're tagging photos in social media. Potentially life-changing when used by law enforcement.
- May mirror human biases and stereotypes, or even amplify them.

[1] This 8min video by an African-American computer scientist illustrates the problem: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

What can cause algorithmic bias?

Among other things, choice of training data.

1. What properties of face recognition training data might cause systems to work poorly on Black faces?
2. What could potentially be done to improve performance on this group?
3. If we did this, might there be other negative consequences?

What can cause algorithmic bias?

Among other things, choice of training data.

- If training examples don't include enough of one type (e.g., Black faces), system won't learn those as well.
- Datasets that match population demographics (or worse, developer demographics) might work well "on average", but not on specific groups.
- Systems developed in East Asia work better on East Asian faces (relative to White) than systems developed in USA.

Could improving performance on Black faces make overall accuracy worse? Isn't it always better to improve overall accuracy?

Well... maybe not

Aside from ethical considerations, it might even be illegal.

- Many countries have laws prohibiting discrimination against certain groups, codifying agreed ethical principles.
- In the UK, the Equality Act of 2010 prohibits discrimination on the basis of nine **protected characteristics**:
  - age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, sexual orientation
  - The law covers many areas, include **services** (free or paid).

Direct and indirect discrimination

Both are illegal.

- **Direct discrimination**: when someone is treated differently to others because they belong to a protected group.
  - An employer who won't hire transgender workers.
  - A pub that only allows women to sit in a particular area.

Direct and indirect discrimination

Both are illegal.

- **Direct discrimination**: when someone is treated differently to others because they belong to a protected group.
  - An employer who won't hire transgender workers.
  - A pub that only allows women to sit in a particular area.
- **Indirect discrimination**: when a policy or practice is applied equally to everyone, but disadvantages people in a protected group more than others.
  - Can you think of examples?

Indirect discrimination

- Example: a shop that requires customers to remove their headgear disadvantages members of some religious groups.
  - This is illegal **indirect discrimination**...
  - ...unless the shop shows the policy is "a proportionate means of achieving a legitimate aim".
  - Say, because of legitimate safety concerns.

Indirect discrimination

- Example: a shop that requires customers to remove their headgear disadvantages members of some religious groups.
  - This is illegal **indirect discrimination**...
  - ...unless the shop shows the policy is "a proportionate means of achieving a legitimate aim".
  - Say, because of legitimate safety concerns.
- I am not a legal expert, so can't say whether the specific AI systems we discuss are illegally discriminating, but
  - Whether legal or not, there can be ethical concerns.
  - If you work at a company you should definitely be considering these questions...

Does NLP disadvantage some groups?

Clearly, yes.

- Across languages: most work is on English or a few other languages. NLP systems limited or unavailable in most languages.
- But often also **within** languages, due to dialect.

What is a dialect?

No clear definition. Famously, "A language is a dialect with an army and navy". But, roughly:

- Varieties of a language, vary according to
  - **region**: Scottish English vs West Country English.
  - **socioeconomic class** or **culture**: Cockney English vs Multicultural London English vs Received Pronunciation.
- Usually mutually intelligible, but differ in pronunciation, lexical items, sometimes morphology and syntax.
- Many countries have a "standard" dialect (General American English, Received Pronunciation), often more prestigious than others.

Who studies dialects?

Among others: **sociolinguists**.

- Language as a social device: to signal identity, achieve social goals.
- How and why do people control social signaling, what is understood (consciously or unconsciously) by others?
- Computational sociolinguistics: studies these questions using computational methods and data sources (e.g. social media).

What's special about social media?

New form of communication, writing that's often closer to speech.

- Often informal, immediate.
- But persistent, and audience not always known.
- Nonstandard spelling and grammar, new lexical items (lol, idk)

Issues for NLP and connection to ethics

Social media text is very different from standard written text. So,

- Tools trained on traditional annotated corpora struggle with these differences.
- Either need to annotate new data, or use **domain adaptation** methods (or some of both).

The further from "standard" language, the worse performance is.

- Marginalized groups often speak less standard dialects. So, potentially more disadvantaged by NLP tools.
- Example: African-American Vernacular English (AAVE)

What is AAVE (or just AAE)?

- Developed originally in Southern USA amongst slave population.
- Now spoken by many (not all!) African Americans across North America.
  - As with many dialects, some speakers have a "stronger" form, others a "weaker" one.
  - Also some differences between regions and urban/rural.
- Characteristic sound pattern (**phonology**) as well as vocabulary and syntax.

(Some) notable characteristics of AAVE

Phonology (often shows up as spelling changes in social media):

- Use of 'd' for GenAmEng 'th' (dis: *this*, dat: *that*)
- Replacing GenAmEng 'ing' with 'in' (walkin: *walking*)

Verb system:

- More tense and aspect distinctions than GenAmEng.
- "be" is often dropped in present tense (she nice: *she's nice*)

Blodgett and O'Connor (2017)

Main question: Do off-the-shelf language ID tools disadvantage African American (AA) Twitter users relative to others?

- i.e., are tools less accurate in predicting "English" on AA language than on "white-aligned" language?
- This could affect who sees the tweets, including downstream applications that pre-filter for English (e.g., sentiment analysis, summarization, etc).

To test this, first need to collect AA and non-AA tweets...

How to collect AA tweets?

- Collect tweets that have **geotag** locations.
- Use US Census data to find racial makeup of neighborhoods.
- Use this data to build a model of AA language (unigram LM!)
- Model can then be applied to any tweet: how AA-like is it?

Modelling dialect variation

Another generative model!

- Assume each word of a tweet is generated as follows:
  - choose which LM to use (four options, including AA and White).
  - then choose a word according to that LM's probability distribution.
- Model is constrained: for each user, the probability of using each LM is similar to their region's racial makeup.
- This allows the model to disentangle which words are most associated with each LM.

The experiment

- Using dialect model, identify "AA-aligned" and "White-aligned" tweets: those where more than 80% of tokens come from AA or Wh LMs.
- Let c be the event that a language ID system correctly IDs the tweet as English.
- For four off-the-shelf systems, compute the difference in accuracy between Wh-aligned and AA-aligned tweets:

$$\text{Diff} = P(c \mid Wh) - P(c \mid AA)$$

Let's consider

Suppose Diff > 0. Can we conclude:

1. The system is less accurate on tweets from AA users?
2. The system is less accurate on tweets that include AA-aligned language?
3. Using AA-aligned language in a tweet causes the system to be less accurate?

Be careful with causation

Just because AA-aligned tweets have lower accuracy doesn't mean AA **caused** lower accuracy.

Could there be another reason? (a **confound**)

Be careful with causation

Just because AA-aligned tweets have lower accuracy doesn't mean AA **caused** lower accuracy.

Could there be another reason? (a **confound**)

- AA-aligned tweets are **shorter** than Wh-aligned tweets.
- We know that language ID is harder for shorter messages.
- So, is the problem AA language or message length?

Results by length (t)

		AA Acc.	WH Acc.	Diff.
langid.py	$t \leq 5$	68.0	70.8	2.8
	$5 < t \leq 10$	84.6	91.6	7.0
	$10 < t \leq 15$	93.0	98.0	5.0
	$t > 15$	96.2	99.8	3.6
IBM Watson	$t \leq 5$	62.8	77.9	15.1
	$5 < t \leq 10$	91.9	95.7	3.8
	$10 < t \leq 15$	96.4	99.0	2.6
	$t > 15$	98.0	99.6	1.6
Microsoft Azure	$t \leq 5$	87.6	94.2	6.6
	$5 < t \leq 10$	98.5	99.6	1.1
	$10 < t \leq 15$	99.6	99.9	0.3
	$t > 15$	99.5	99.9	0.4
Twitter	$t \leq 5$	54.0	73.7	19.7
	$5 < t \leq 10$	87.5	91.5	4.0
	$10 < t \leq 15$	95.7	96.0	0.3
	$t > 15$	98.5	95.1	-3.0

Conclusions

- (As expected) accuracy is lower on shorter tweets.
- For nearly all systems and lengths, accuracy is higher for Wh-aligned tweets than AA-aligned tweets.
- The difference is particularly large for the shortest tweets.

Is the difference "meaningful"?

Statistically significant?

- Is it possible these differences are simply due to random chance?
  - We'll discuss this concept later in the course; but no, the authors show their results are not just random (footnote 11).

Is the difference "meaningful"?

Statistically significant?

- Is it possible these differences are simply due to random chance?
  - We'll discuss this concept later in the course; but no, the authors show their results are not just random (footnote 11).

Large enough to matter?

- What information that I haven't told you might help decide?

Is the difference "meaningful"?

Statistically significant?

- Is it possible these differences are simply due to random chance?
  - We'll discuss this concept later in the course; but no, the authors show their results are not just random (footnote 11).

Large enough to matter?

- What information that I haven't told you might help decide?
  - Maybe lots, but at least: what proportion of AA tweets are $t \leq 5$? (Answer: over 40%)
  - That's a **lot** of tweets. So downstream applications could strongly underrepresent the data from African Americans.

Are there solutions?

No easy ones, but various starting points:

- Adapting existing models to better handle specific dialects, or building dialect-sensitive models from the start (as in this paper!)
- Developing (and using!) methods for testing algorithmic bias.
- Considering possible problems **before** building a system rather than trying to retrofit afterwards.

Awareness of these issues is rising but by no means universal.

- As usual, much of the discussion centers around English. Maybe some of you will help broaden that discussion to other languages.

Summary

- Language use varies to express social factors (formality, identity).
- Variability is especially widespread in user-created text, but NLP systems have trouble dealing with it.
- This can create unintentional disadvantages for some groups.
  - We'll see other examples of algorithmic bias in NLP later on.
- Blodgett and O'Connor study: quantifies racial disparity in language ID; good example of careful experiment design/analysis.
- Both areas (algorithmic bias; language variation) are growing research topics in NLP.

Questions and exercises

- Why is the dialect model constrained at the user level rather than the tweet level? That is, why not say that each tweet (rather than each user) uses the AAVE LM with probability similar to the region's racial makeup?
- Algorithmic bias can arise if the training data is not representative, or doesn't include enough examples from some category. But it can also arise if the annotations are themselves biased (i.e., have more errors on one category than another). Suppose you want to annotate a data set that contains a lot of AAVE, but most of your annotators are white. What problems could this cause with the annotation and with the resulting NLP system?

References

Blodgett, S. L. and O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, KDD*.