Clickbait Detection using Natural Language Processing and Machine Learning

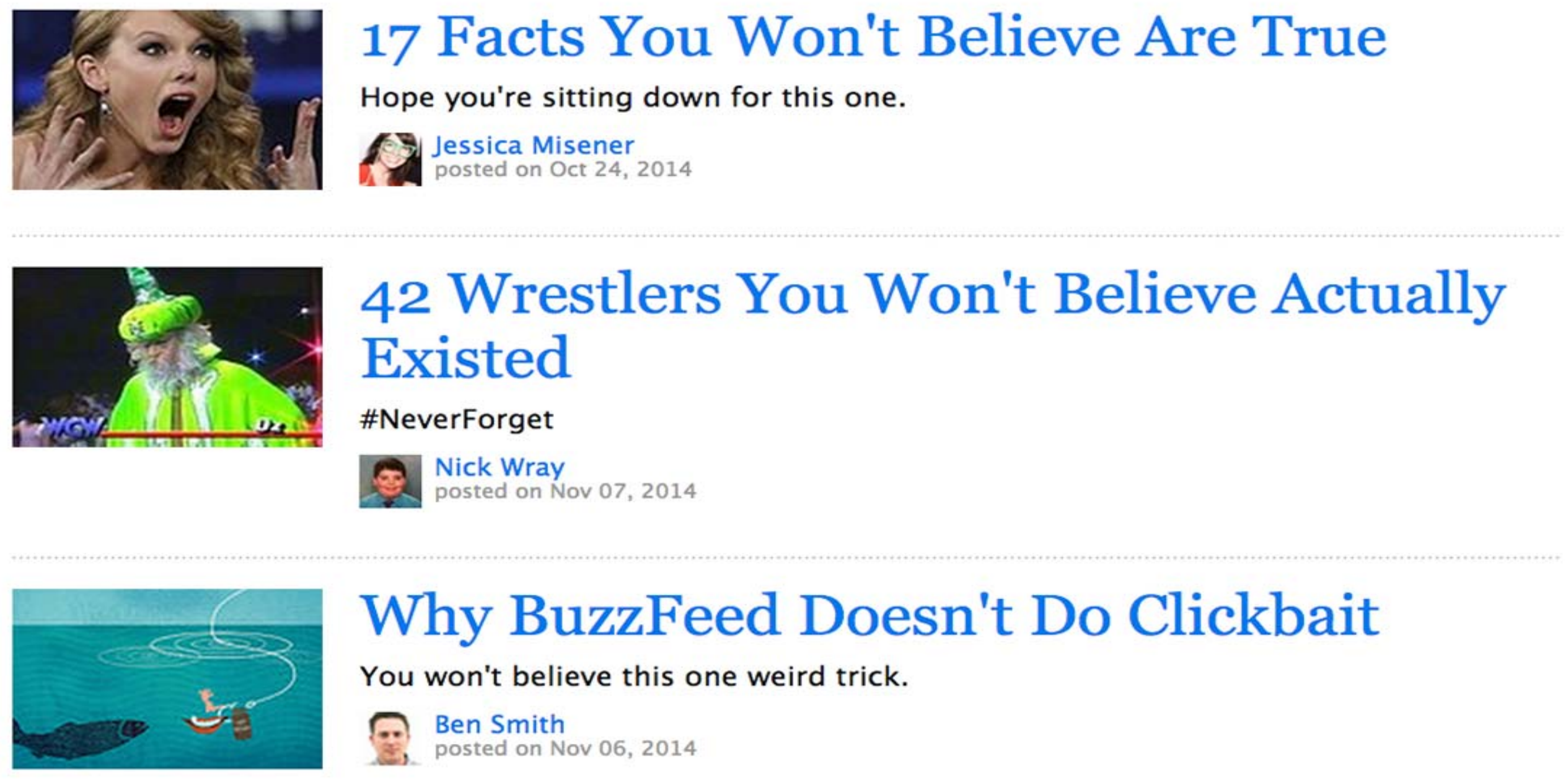
UNION COLLEGE

Varun Shah

Kristina Striegnitz and Nick Webb, Advisors

What is Clickbait?

- Social Media posts designed to entice the clicking of an accompanying link in order to increase online readership.
- Clickbait usage by news publishers could give rise to echo chambers of false information and fake news.



17 Facts You Won't Believe Are True

Hope you're sitting down for this one.

Jessica Misener
posted on Oct 24, 2014

42 Wrestlers You Won't Believe Actually Existed

#NeverForget

Nick Wray
posted on Nov 07, 2014

Why BuzzFeed Doesn't Do Clickbait

You won't believe this one weird trick.

Ben Smith
posted on Nov 06, 2014

Figure 1. Examples of Clickbait. Source: *www.baekdal.com*

Research Question

- How can one determine whether a post on social media is clickbait?
- Utilize Natural Language Processing and Machine Learning in order to develop a model that accurately predicts Clickbaiting.
- Study and understand what makes a post Clickbait or not and analyze how well our classifier can detect it.

The Data

- We use the *clickbait17-train* datasets[1] with 2451 instances and the following important attributes:
  - *postText:* Text of the post without the link
  - *targetTitle:* Title of the target article
  - *truthClass:* Whether post is clickbait or no-clickbait
- Examples of clickbait:
  - *What India's microloan meltdown taught one entrepreneur*
  - *31 Accessories Every 90s Girl Will Recognize*.
- Examples of no-clickbait:
  -Prince Harry meets Lady Gaga at the Royal Albert Hall
  -Apple debuts iOS 9: Battery enhancements smarter Siri.

Methods

- Resampled data to obtain uniformly distributed class attribute.
- Attributes included in the model were *postText*, *targetTitle*, and *truthClass.*
- Ran several 10-fold cross-validation classifications on the data experimenting with various classification algorithms like *ZeroR*, *J48*, *LibSVM*, and *RandomForest* in order to determine which can most accurately detect clickbait.

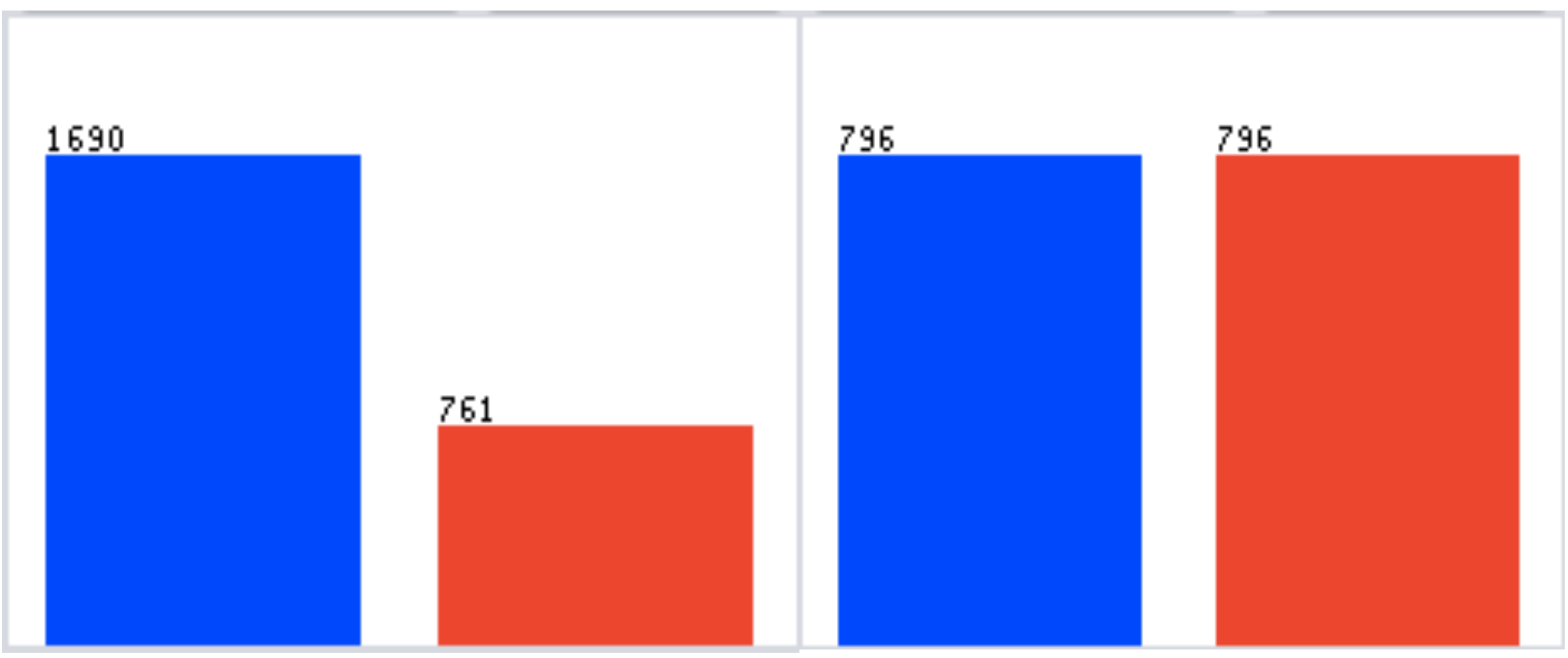


Figure 2. Distribution of class attribute before and after resampling

Preliminary Results

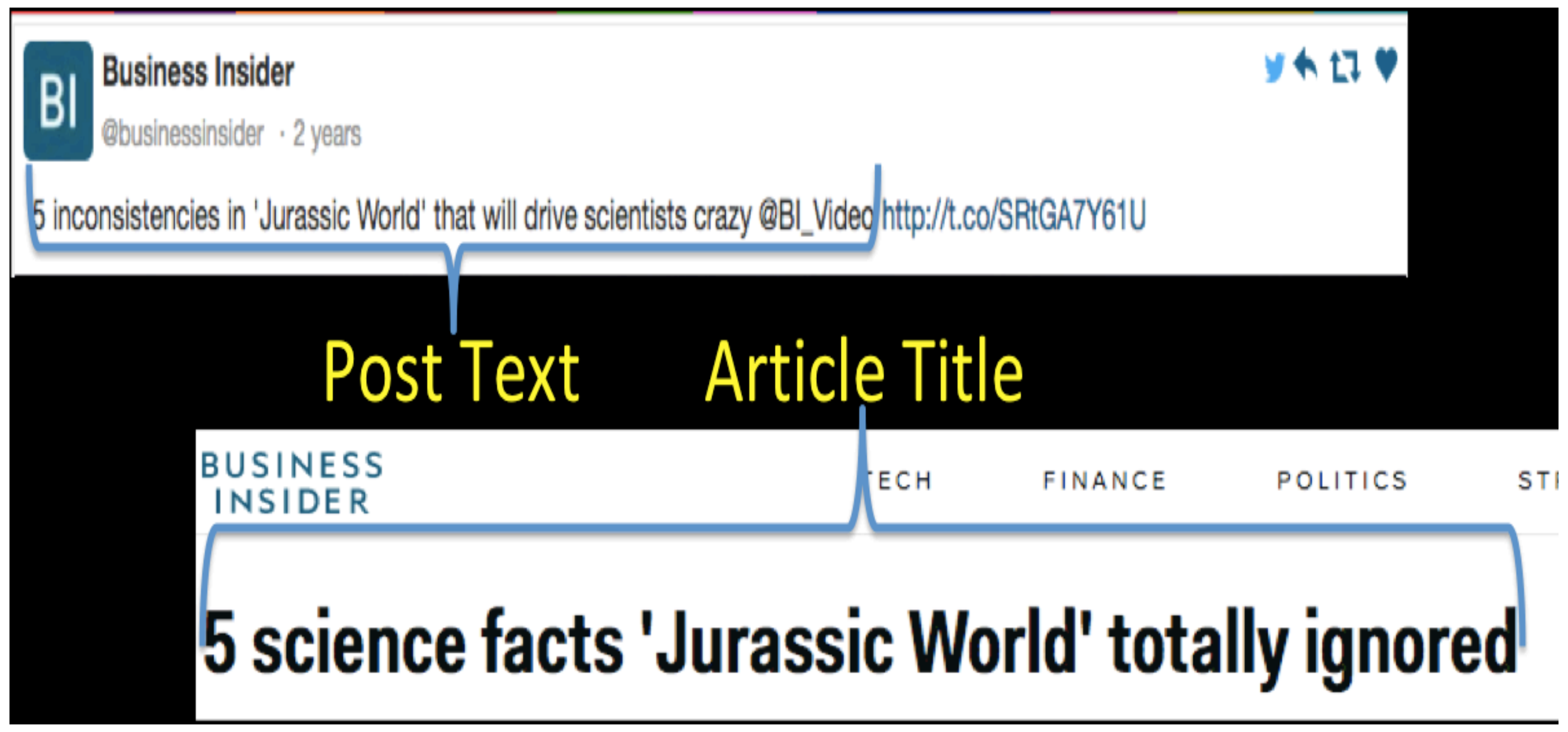
- Obtained the following statistically significant results:

Classifier	Classification Accuracy
ZeroR	50.0%
J48	76.3819%
LibSVM	82.1608%
RandomForest	86.3065%

- *ZeroR* is our baseline prediction algorithm and always chooses the majority class. We use it as a reference point to evaluate the performance of other classifying algorithms.
- RandomForest achieves highest statistically significant accuracy and so becomes our classification algorithm.

Added Features

- Added 25 features out of which 3 worked:
- ***numWords***: The number of words in *postText*. Lower the number of words, the more likely a post is clickbait.
- ***numOverTitle***: The number of overlapping words between postText and targetTitle. Higher the number of overlaps, the more likely a post is clickbait.
- ***posRatio***: The likelihood that a parts-of-speech sequence appears in clickbait instances. Higher the likelihood of a POS sequence appearing in clickbait, the more likely a post is clickbait.



Business Insider
@businessinsider · 2 years

5 inconsistencies in 'Jurassic World' that will drive scientists crazy @BI_Video http://t.co/SRtGA7Y61U

5 science facts 'Jurassic World' totally ignored

Figure 3. Example of *postText* and *targetTitle*

POS Ratio = #Sequence in Clickbait / #Sequence in All

Figure 4. posRatio formula

Results and Conclusion

- Results obtained by adding features to the model and performing a 10-fold cross-validation on unseen data:

Attributes	Accuracy
postText + targetTitle + numWords + numOverTitle	82.2864%
postText + targetTitle + posRatio	86.6860%
postText + targetTitle + numWords + numOverTitle + posRatio	88.2051%

- We conclude that our model is good at detecting clickbait, and that the number of words in the postText, similarity between the postText and targetTitle, as well as the Parts-of-speech ratio are useful features in clickbait detection.

Acknowledgments

I would like to thank my project advisors, Prof. Kristina Striegnitz and Prof. Nick Webb, as well as my academic advisor, Prof. Fernandes.

References

[1] The Clickbait Challenge 2017. http://www.clickbait-challenge.org/

[2] Business Intelligence. *Data Mining with R: J48 decision tree.*

[3] Chih-Chung Chang, Chih-Jen Lin. *LIBSVM – A library for Support Vector Machines.*