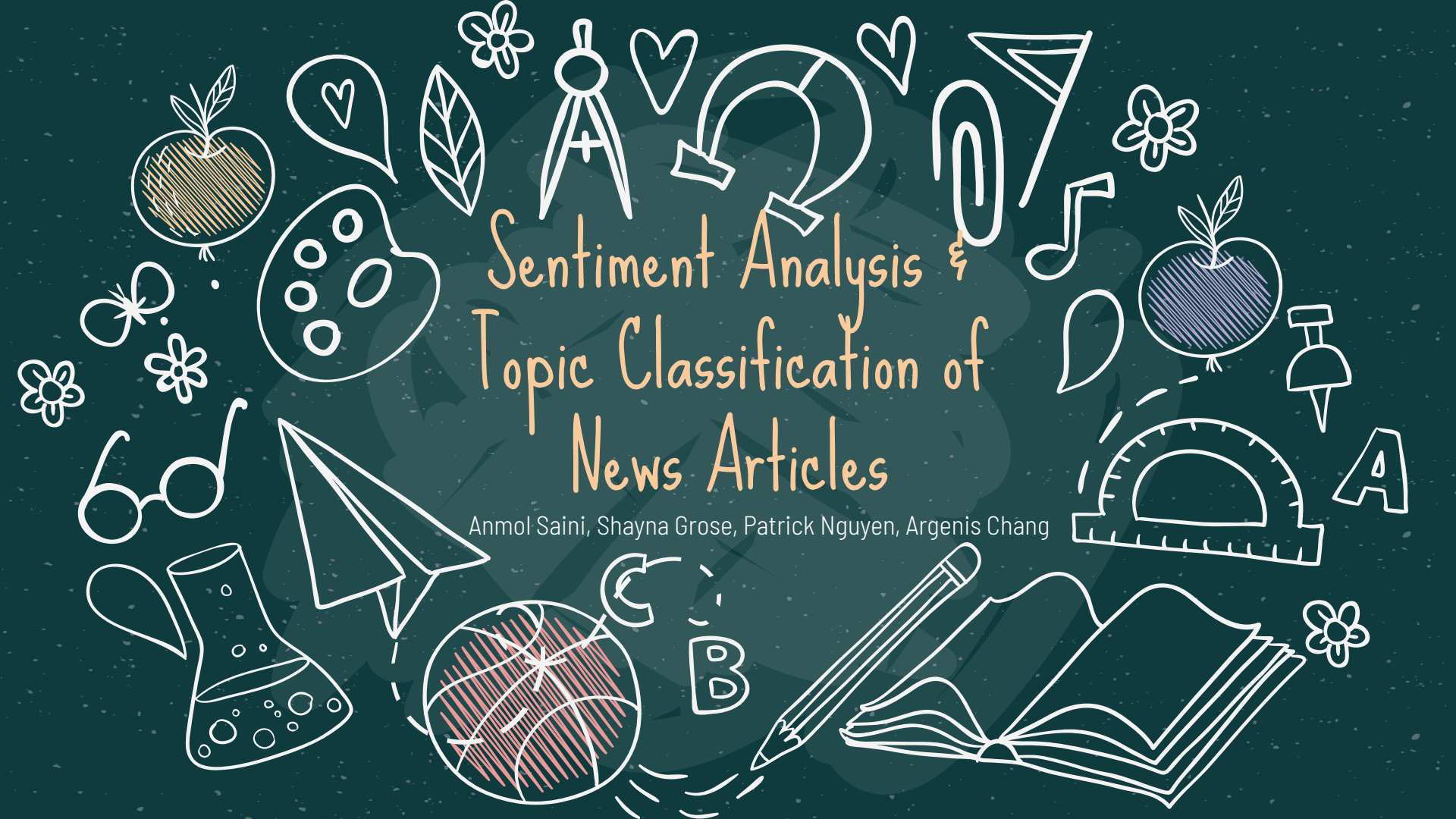

Sentiment Analysis &
Topic Classification of
News Articles
Anmol Saini, Shayna Grose, Patrick Nguyen, Argenis Chang
A
B
C

PURPOSE

- Most news media is consumed online and with the wealth of information online, newspaper outlets have turned to sensationalizing and polarizing news articles headlines to grab readers attention.
- We want to classify articles into positive and negative categories based on just their contents.
- We will also perform topic classification of the news articles, to see if such sensationalizing happens more often in certain genres of news than others.

OUR VISION

When we started this project we envisioned a map based user interface where users could see news happening around the world, and filter what to read based on location, news category, and the sentiment of the news itself.

01

Introduction

Our datasets and how we found them.

02

Sentiment Analysis

The different sentiment analysis methods we tried.

03

Topic Classification

How NMF and SVC works.

04

Future Work

Where we could go from here.

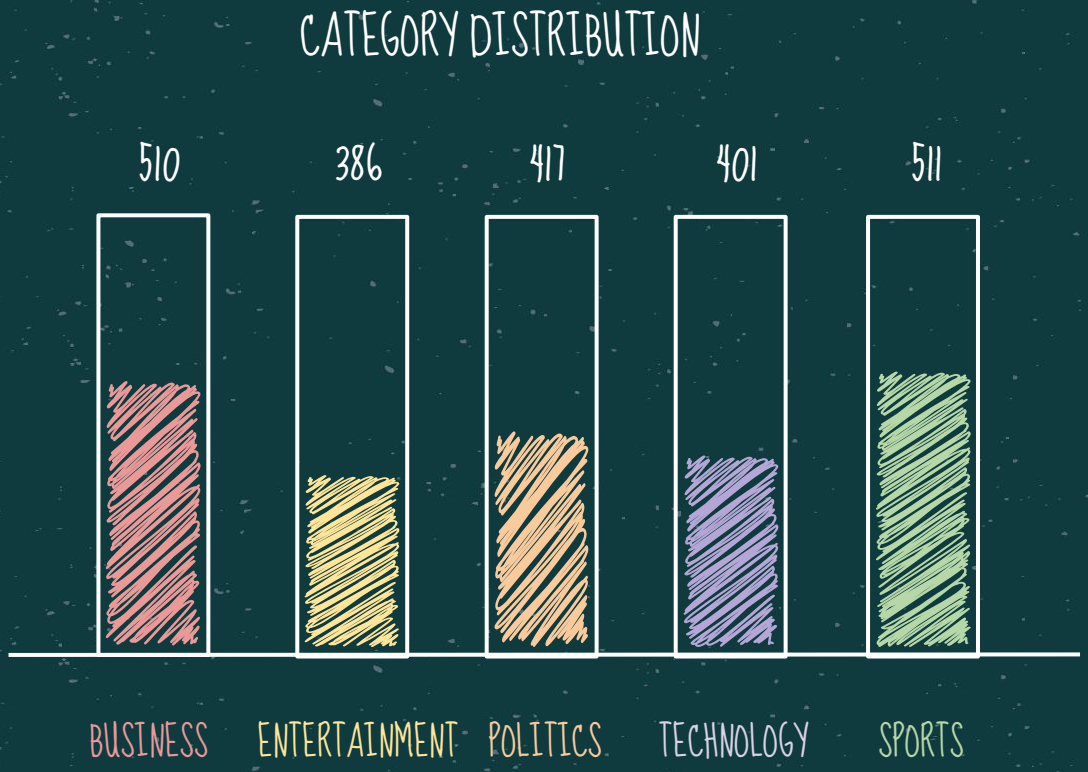
1. INTRODUCTION

DATA SETS

We used a British Broadcasting Company dataset of 2225 newspaper articles from the years 2004-2005, that were categorized into 5 categories:

- Business
- Entertainment
- Sports
- Politics
- Technology

Each article contained a title and the contents of the article, with an average article length of 379 words


CATEGORY DISTRIBUTION
510
386
417
401
511
BUSINESS
ENTERTAINMENT
POLITICS
TECHNOLOGY
SPORTS

2. SENTIMENT ANALYSIS

WE CONDUCTED THE POSITIVE AND NEGATIVE CLASSIFICATION USING TWO METHODS

RAW COUNTS

This method counted the total number of positive and negative words in an article, and whichever was greater, that would be the classification of the article.

TF-IDF

We trained a TF-IDF model on all the article contents, and then used that model to find the words in each article that are “important”, and then count the number of positive and negative ones (similar to our other method.)

STEPS IN FINDING THE COUNTS

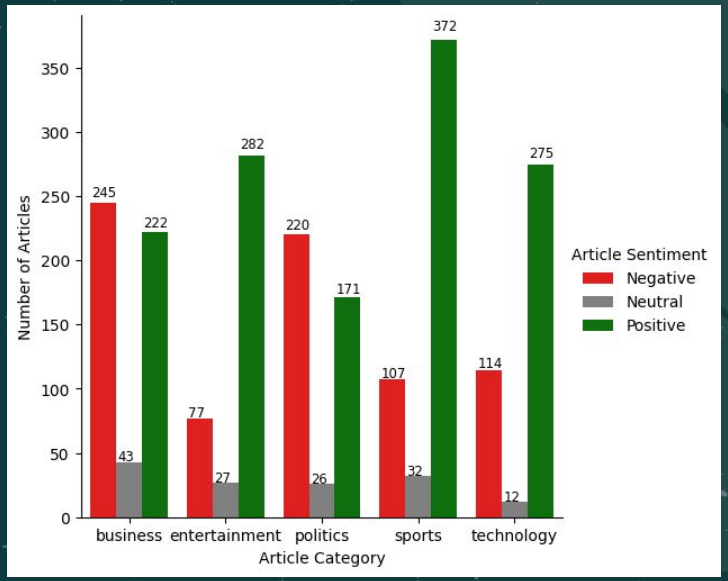
01 We used a positive and negative word corpus containing 4783 negative words and 2007 positive words.

02 We then compared each word in the articles contents to these corpuses to see if any matched.

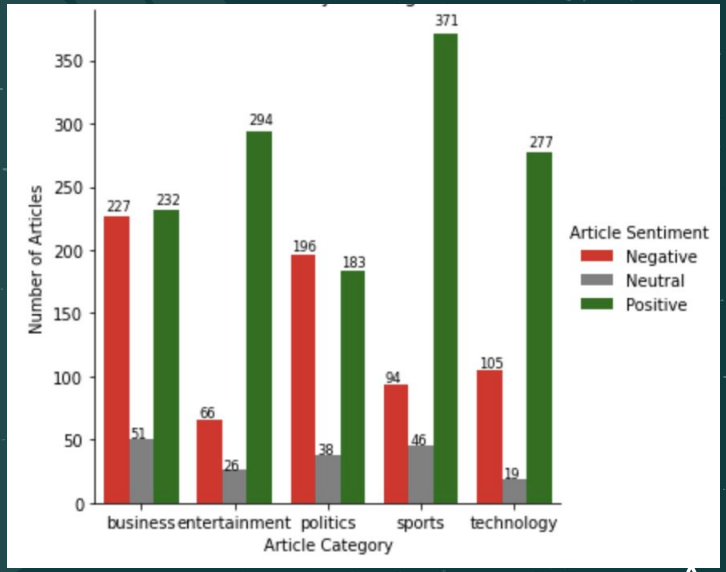
03 For every matching positive or negative word, we increased a counter, and then at the end we compared these two totals, and classified based on the greater number.

04 We also repeated this process, but just using article titles to see if the results were different.

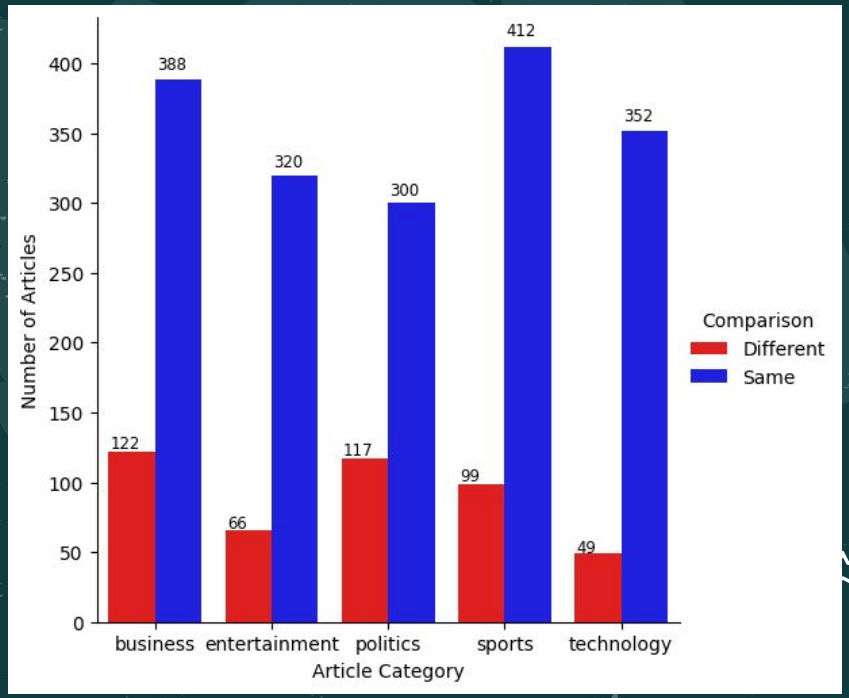
We then repeated this whole process, but used only the “important” words in an article found by TF-IDF



Sentiment classification of articles using raw counts.



Sentiment classification of articles using TF-IDF



Comparison of the results from raw count classification and tf-idf sentiment classification.

OUR FINDINGS

1

TITLES vs. CONTENTS

We found that trying to determine the sentiment using just an articles title mostly returned "neutral" since titles often didn't have any positive or negative words.

2

TF-IDF vs. RAW COUNTS

Since the sentiment of the articles is unlabelled we can't compare accuracy, but for each category over half of the articles were classified the same by both methods.

3

TRENDS

In both methods, we found that the "Politics" category was the one that tended to have more negative than positive news articles.

3. TOPIC CLASSIFICATION

OUR TOPIC CLASSIFICATION PIPELINE

01

TEXT PROCESSING

We employed a number of different cleaning methods including punctuation removal, stemming, stop word removal, and contraction expansion.

02

NMF

We used Non-Negative Matrix Factorization on the TF-IDF term frequency matrices for articles to find a topic for each article consisting of the 8 best words to describe it.

03

SVC

We also trained a Support Vector Classifier on the TF-IDF term frequency matrices for articles, and their labelled categories.

04

CLASSIFICATION

Together, NMF gives each article 8 keywords as a title, and the SVC maps the titles given to one of our 5 original article categories.

EXAMPLE


1
ARTICLE
"Hollywood stars Kevin Spacey and Kate Bosworth attended the British premiere of new film, Beyond the Sea, in London's Leicester Square on Thursday..."
2
NMF GENERATED TOPIC
Film, festive, star, movies, director, actor, cinema, oscars
3
SVC PREDICTED CATEGORY
Entertainment

RESULTS

Category	Correctly classified articles	Accuracy
Business	418	83.43%
Entertainment	346	94.79%
Politics	323	81.36%
Sports	496	99.20%
Technology	135	40.54%

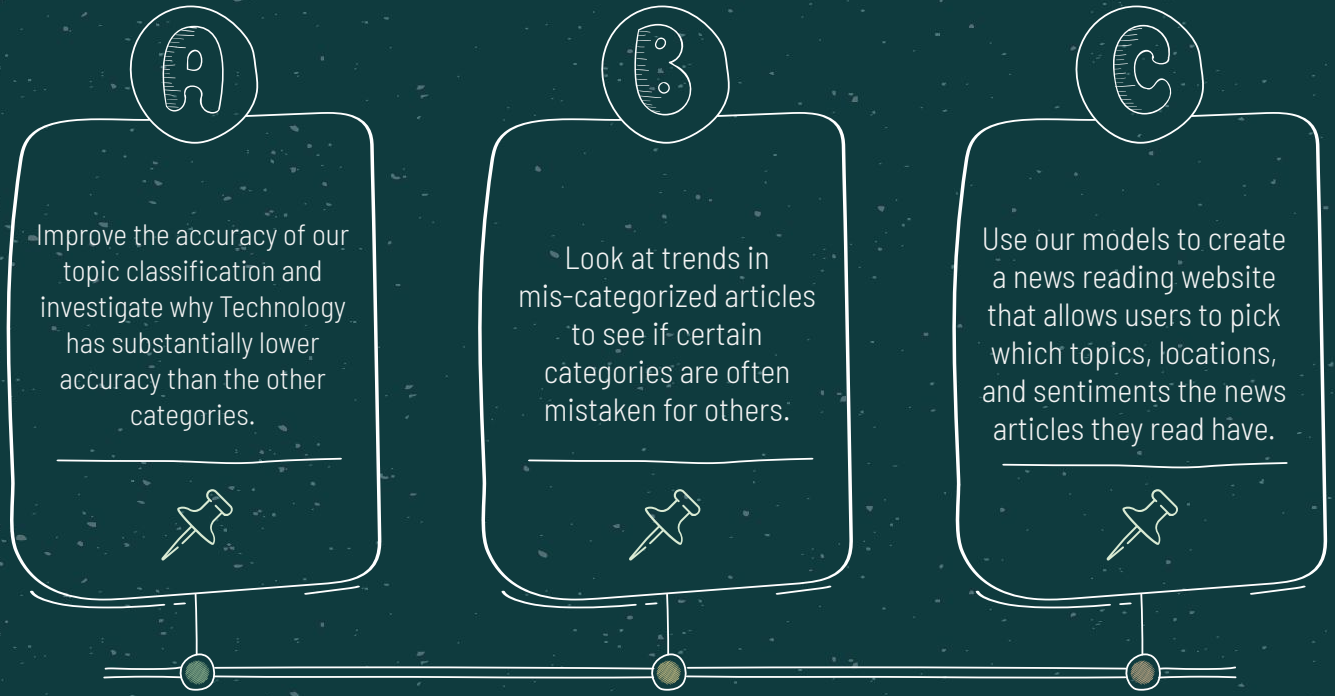
MAIN FINDINGS

We found that:

- Sports articles were most often correctly categorized by our models
- Technology had the lowest correct classification rate
  - We think this may be due to the fact that there is lots of overlap potentially between sports and technology (ie. words like game, play etc), and entertainment and technology
- We tested NMF using different numbers of topic words, and topics to generate, and found that 15 topics using 8 keywords produced the best results

4. FUTURE WORK

FUTURE WORK


A
Improve the accuracy of our topic classification and investigate why Technology has substantially lower accuracy than the other categories.
B
Look at trends in mis-categorized articles to see if certain categories are often mistaken for others.
C
Use our models to create a news reading website that allows users to pick which topics, locations, and sentiments the news articles they read have.

THANKS!

Shayna Grose
Anmol Saini
Patrick Nguyen
Argenis Chang

REFERENCES

1. Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

2. Neal Shyam. 2014. Expand common english contractions. *GitHub repository*

3. Soonh Taj, Baby Bakhtawer Shaikh, and Areej Fatemah Meghji. 2019. Sentiment analysis of news articles: A lexicon based approach. 2019 International Conference on Computing, Mathematics and Engineering Technologies

4. D. Greene and P. Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering.International Conference on Machine Learning.