An Article Classifier

Know Before You Read

By Carla and Vincent

1

Background

Problem Statement & Objective

2

Methodology & Results

Data & Process

Results & Limitations

3

Recommendations

For Practitioners & Stakeholders

1. Background

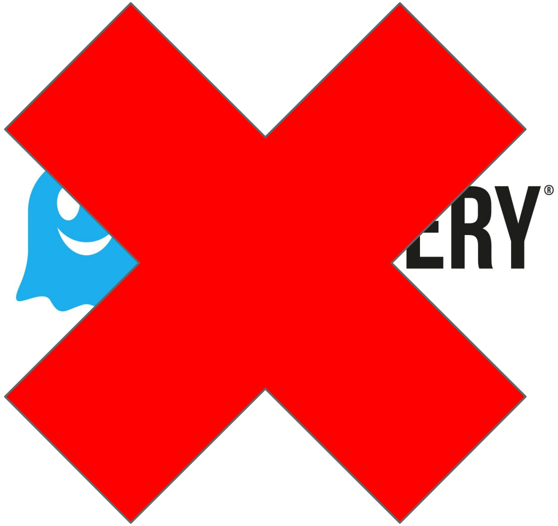
Problem Statement & Objective

Problem Statement

The reasoning behind our problem...

Information Overload

Absence of Tool in the Market



Current Research Without Business Implications

Detecting Promotional Content in Wikipedia

Shruti Bhosale **Heath Vinicombe** **Raymond J. Mooney**
Department of Computer Science
The University of Texas at Austin
{shruti,vini,mooney}@cs.utexas.edu

Abstract

This paper presents an approach for detecting promotional content in Wikipedia. By incorporating stylometric features, including features based on n-gram and PCFG language models, we demonstrate improved accuracy at identifying promotional articles, compared to using only lexical information and meta-features.

based on both n-grams and Probabilistic Context Free Grammars (PCFGs). We show that using such stylometric features improves over using only shallow lexical and meta-features.

2 Related Work

Anderka et al. (2012) developed a general model for detecting ten of Wikipedia's most frequent quality flaws. One of these flaw types, "Advert"[2], refers to

Bert-Based Promotional Words Detection Classifier Development

Zikun Lin (supervided by: John R. Kender)

Columbia University, New York, NY, United States

Abstract

From the analysis last semester, we can see the necessity for us to find a way to get rid of advertising and promotion words from our corpora. After manually picking out promotion sentences, we get a list of "blackwords" from video descriptions corpora, which can be used as a dataset in machine learning. In this report, I use the traditional model Word2Vec and the most advanced NLP model "BERT" combining with other machine learning methods to help us picking out the promotion sentences from video descriptions.

Objective

“Provide a convenient method in article classification for web browsers such as Chrome, Firefox, or Safari”.

Key Goals

- Build a classifier for promotional articles
- Improve web-browsers capabilities
- Apply text analytics techniques
- Provide a new solution to an existing problem

2. Methodology & Results

Data Information, Process, Results
& Limitations

Data Information

Text Data
50,000 total
2 Datasets

kaggle

Promotional.csv

Labels: advert, coi, fanpov, pr & resume

Good.csv

Articles classified as "good articles"

```
[ ] good.head()
```

	text	url
0	Nycticebus linglom is a fossil strepsirrhine p...	https://en.wikipedia.org/wiki/%3F%20Nycticebus...
1	Oryzomys pliocaenicus is a fossil rodent from ...	https://en.wikipedia.org/wiki/%3F%20Oryzomys%2...
2	.hack dt hk is a series of single player actio...	https://en.wikipedia.org/wiki/.hack%20%28video...
3	The You Drive Me Crazy Tour was the second con...	https://en.wikipedia.org/wiki/%28You%20Drive%2...
4	0 8 4 is the second episode of the first seaso...	https://en.wikipedia.org/wiki/0-8-4

```
[ ] good.describe()
```

	text	url
count	30279	30279
unique	30279	30279
top	Hurricane Joanne was one of four tropical cycl...	https://en.wikipedia.org/wiki/Bernard%20Waldman
freq	1	1

```
[ ] promotional.head()
```

	text	advert	coi	fanpov	pr	resume	url
0	1 Litre no Namida 1, lit. 1 Litre of Tears als...	0	0	1	0	0	https://en.wikipedia.org/wiki/1%20Litre%20no%2...
1	1DayLater was free, web based software that wa...	1	1	0	0	0	https://en.wikipedia.org/wiki/1DayLater
2	1E is a privately owned IT software and servic...	1	0	0	0	0	https://en.wikipedia.org/wiki/1E
3	1Malaysia pronounced One Malaysia in English a...	1	0	0	0	0	https://en.wikipedia.org/wiki/1Malaysia
4	The Jerusalem Biennale, as stated on the Bienn...	1	0	0	0	0	https://en.wikipedia.org/wiki/1st%20Jerusalem%...

```
[ ] promotional.describe()
```

	advert	coi	fanpov	pr	resume
count	23837.000000	23837.000000	23837.000000	23837.000000	23837.000000
mean	0.793346	0.089860	0.062760	0.063599	0.092210
std	0.404913	0.285988	0.242535	0.244042	0.289328
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

Our Process

Data Cleaning & Exploration

- Data Collection
- Data Preparation

Machine Learning Models

- Sentiment Analysis
- Topic Modelling
- Classification with RF

Results

Data Exploration & Cleaning

Describe Our Data

```
[8] promotional.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23837 entries, 0 to 23836
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    23837 non-null  object
 1   advert  23837 non-null  int64
 2   coi     23837 non-null  int64
 3   fanpov  23837 non-null  int64
 4   pr      23837 non-null  int64
 5   resume  23837 non-null  int64
 6   url     23837 non-null  object
dtypes: int64(5), object(2)
memory usage: 1.3+ MB
```

```
good.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30279 entries, 0 to 30278
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    30279 non-null  object
 1   url     30279 non-null  object
dtypes: object(2)
memory usage: 473.2+ KB
```

Check for Null Values

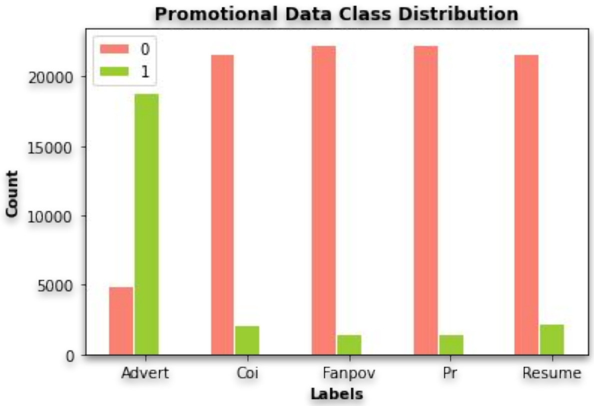
```
[ ] promotional.isnull().sum()

text      0
advert    0
coi       0
fanpov    0
pr        0
resume    0
url       0
dtype: int64
```

```
[ ] good.isnull().sum()

text    0
url     0
dtype: int64
```

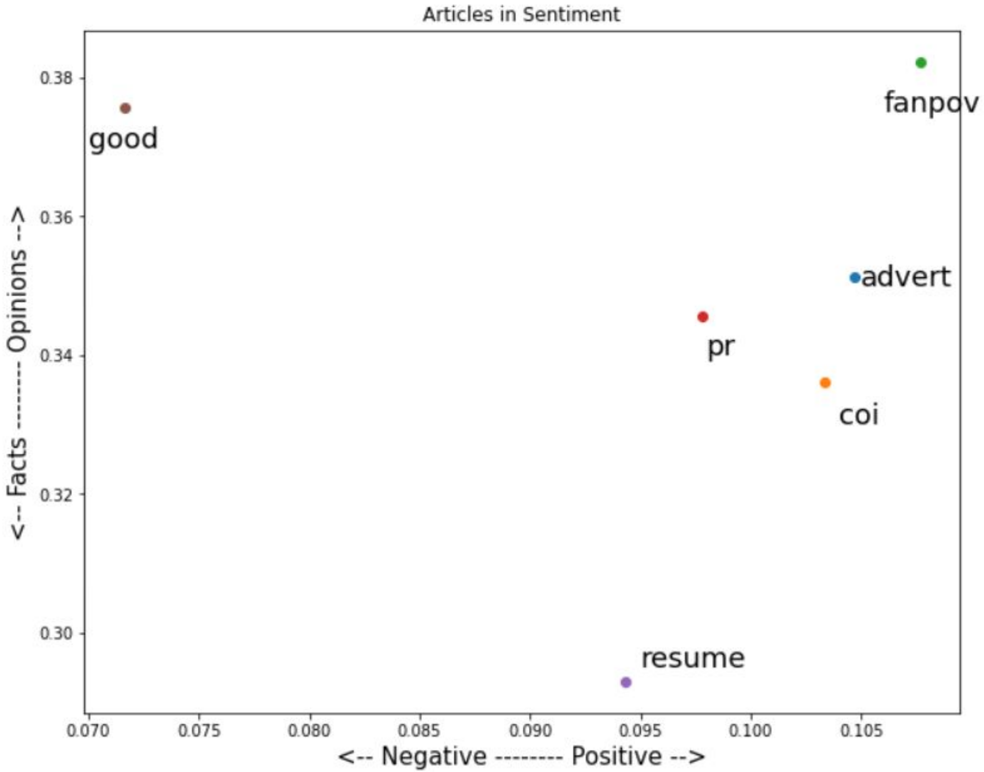
Check Class Distribution



Machine Learning Models

Sentiment Analysis, Topic Modelling & Classification

Sentiment Analysis

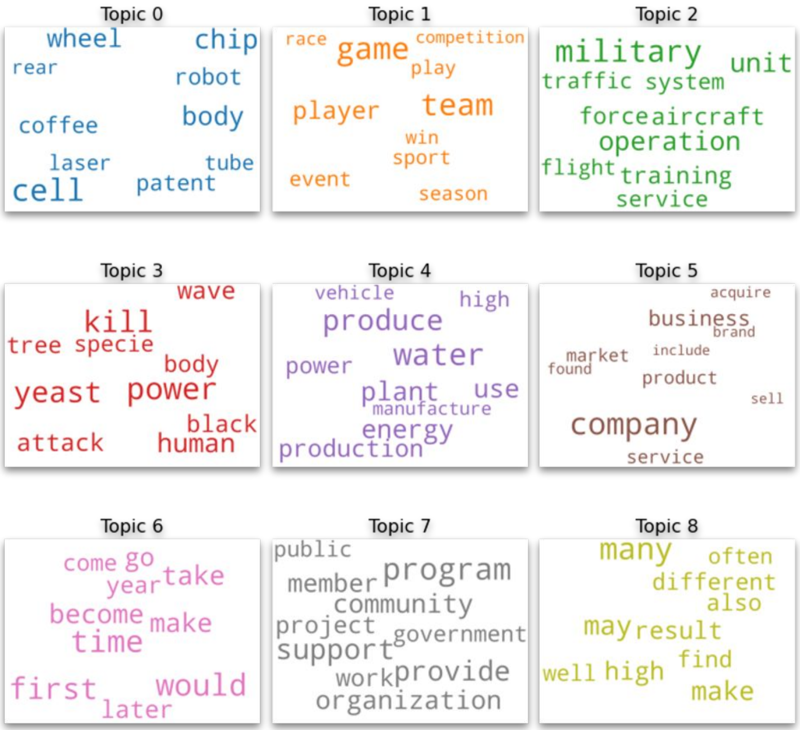


- Good articles usually are thought to be more fact-based and neutral in wording
- Promotional articles are more positive
- Resume-like articles are more negative

Topic Modeling Process

1. Tokenization of texts
2. Remove Stopwords, Create Bigrams & Lemmatization
3. LDA Model
4. Coherence score
5. Wordcloud Visualizations

Promotional Articles



Topic 0: Science/Robotic

Topic 1: Sports

Topic 2: Military

Topic 3: History

Topic 4: Power Supplies

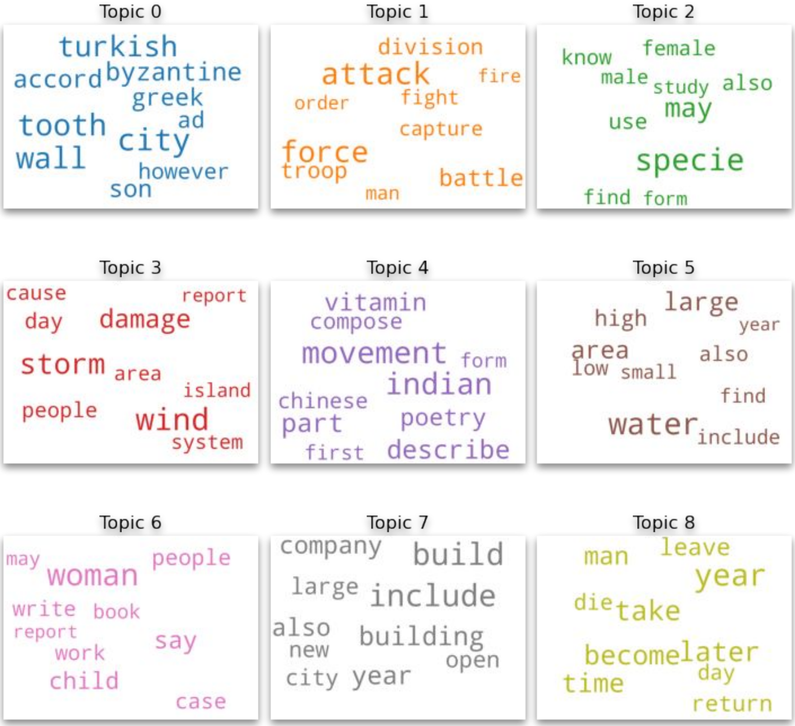
Topic 5: Business

Topic 6: Releases

Topic 7: Society

Topic 8: Results

Good Articles



Topic 0: Culture

Topic 1: War

Topic 2: Science

Topic 3: Weather Conditions

Topic 4: Cultural Movements

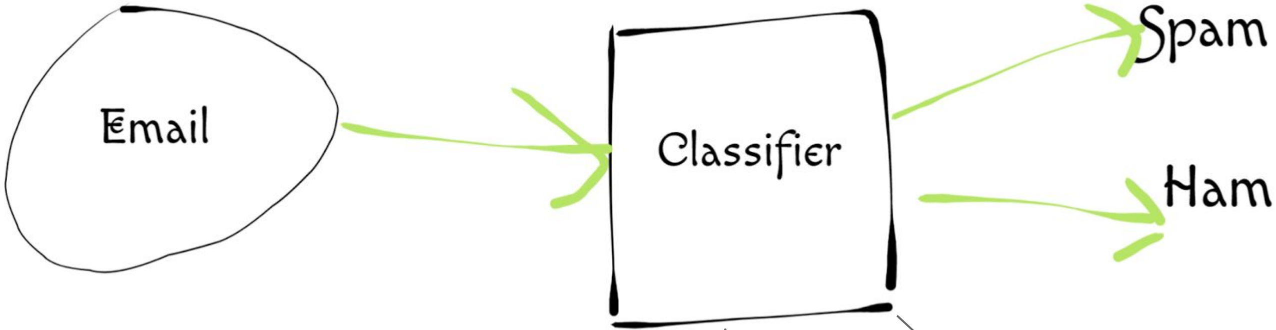
Topic 5: Lands

Topic 6: Books

Topic 7: Companies

Topic 8: Historical Events

Classifier



CountVectorizer

TfidfTransformer

Random forest

Advantages

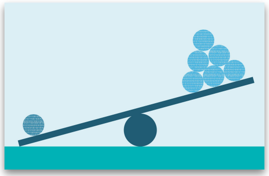
Convenient

Easy Input
Useful Classifications

Accurate

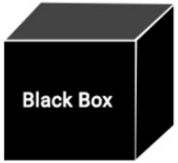
Accuracy Score: 0.81
Resume: Advertisement

Limitations



Imbalanced Data

Most Promotional are Advertisements



Black-Box Model

Which Decision Tree?
How to Tune?

3. Recommendations & Implications

For Practitioners & Stakeholders

CEO of Web Browsers / Director of Quality

- Complexity of articles on the web
- Create a web-browser extension
- Increase customer satisfaction/product quality
- Attract new users

- Improve interest in text analytics
- Develop attractive articles
- Realize people's need of useful information

Marketing Analysts

Websites' Owners / Administrators

- Avoid using articles that are misleading
- Implement this tool in their own articles
- Recognize user's real needs

Credits

Team Members: Vincent Chen, Carla M. Fera

Prof. Nohel Zaman

All faculty members & classmates!!

Thanks!

Does anyone have any questions?