“ALEXANDRU IOAN CUZA” UNIVERSITATY OF IAŞI
FACULTY OF COMPUTER SCIENCE

The Semantics and Pragmatics of Natural Language

Daniela GÎFU

http://profs.info.uaic.ro/~daniela.gifu/

Course 13

Binary Classification.

Techniques to detect propaganda in news

What is this lecture about?

- Meaning binary classification
- Building classifications system
- Application in media
- Conclusion

Binary Classification

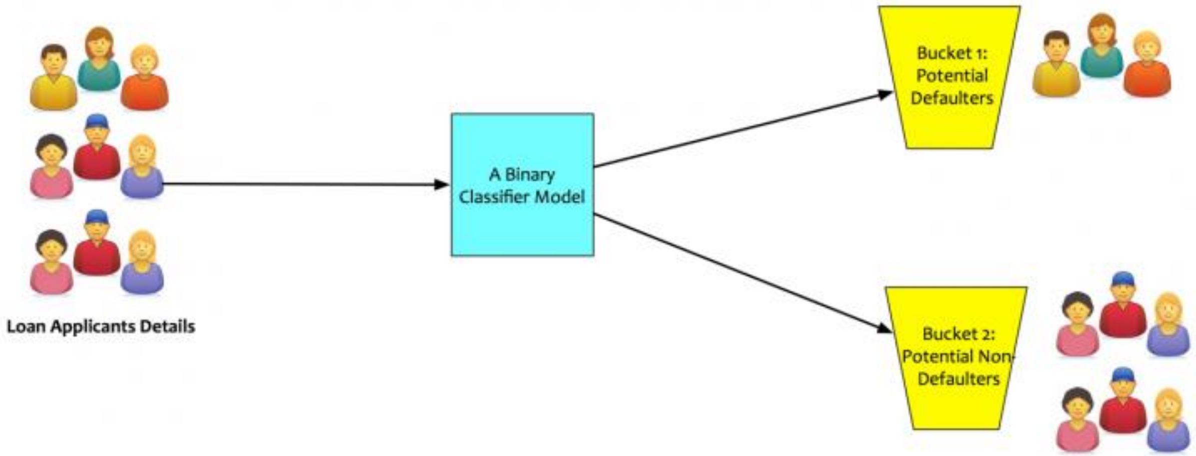
- the task of classifying the elements of a set into **two groups** on the basis of a **classification rule**.

- refers to those **classification tasks** that have **two class labels/** two possible outcomes.

Ex: Email spam detection (spam or not) or Gender classification (Male / Female)

How to build a classification model

- **Initialize** the classifier to be used.
- **Train** the classifier: All classifiers in scikit-learn uses a fit(X, y) method to fit the model(training) for the given train data X and train label y.
- **Predict** the target: Given an unlabeled observation X, the predict(X) returns the predicted label y.
- **Evaluate** the classifier mo

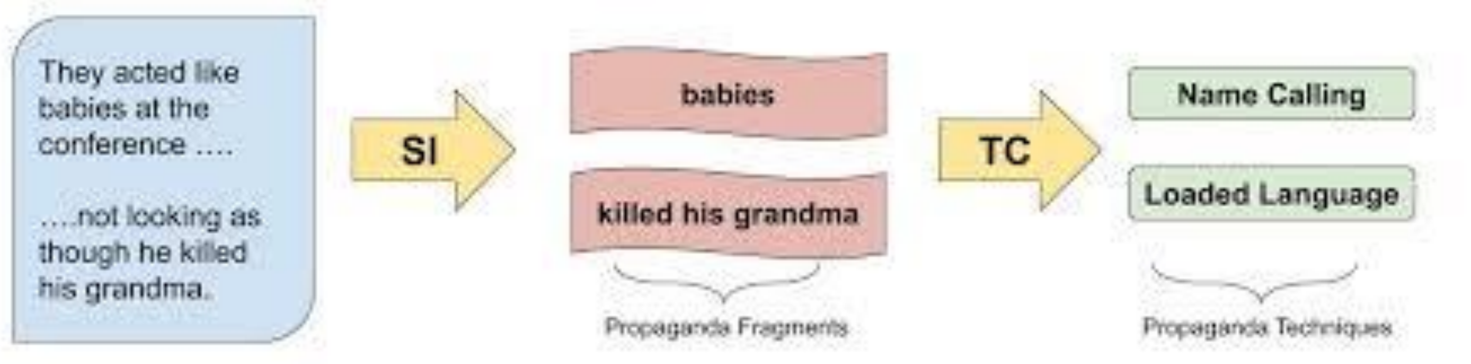

Loan Applicants Details
A Binary Classifier Model
Bucket 1: Potential Defaulters
Bucket 2: Potential Non-Defaulters

Binary Classification. Usecase

Propaganda classification using news articles

Objectives

- Given a plain-text document, to identify those specific sentences which contain at least one propaganda technique.
- Given a sentence classified as propaganda, to identify the applied propaganda technique in the fragment.

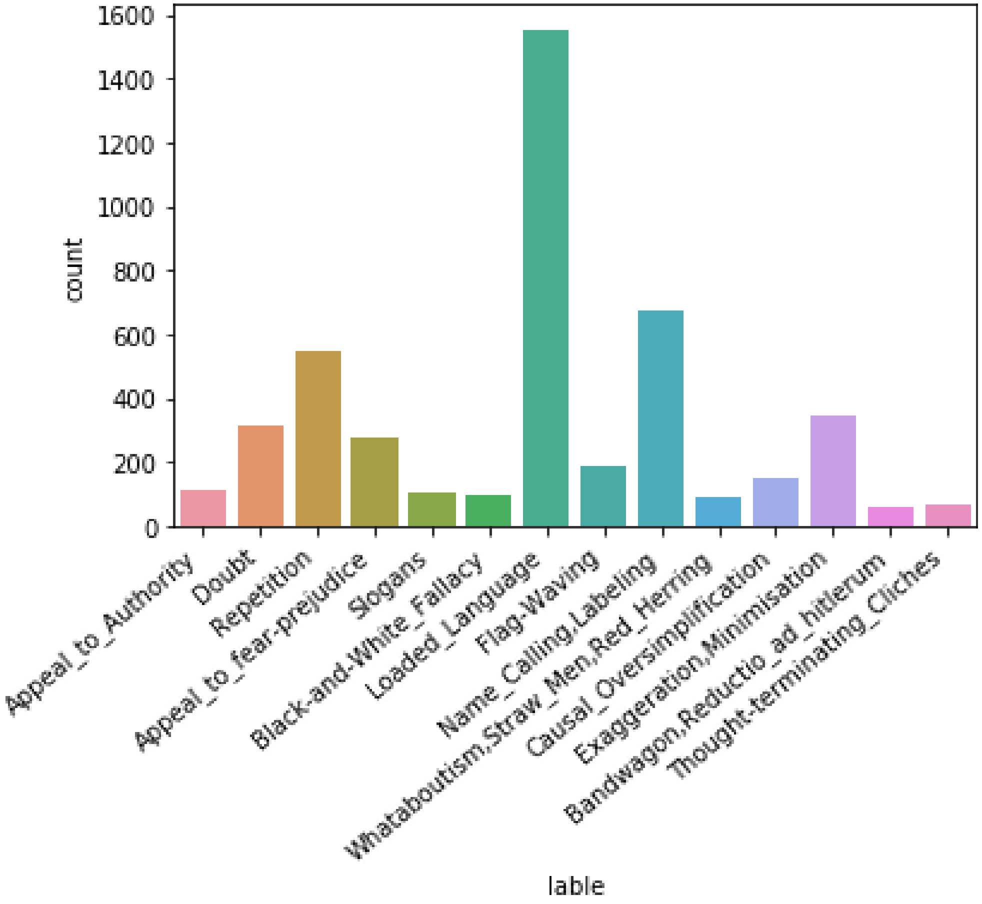


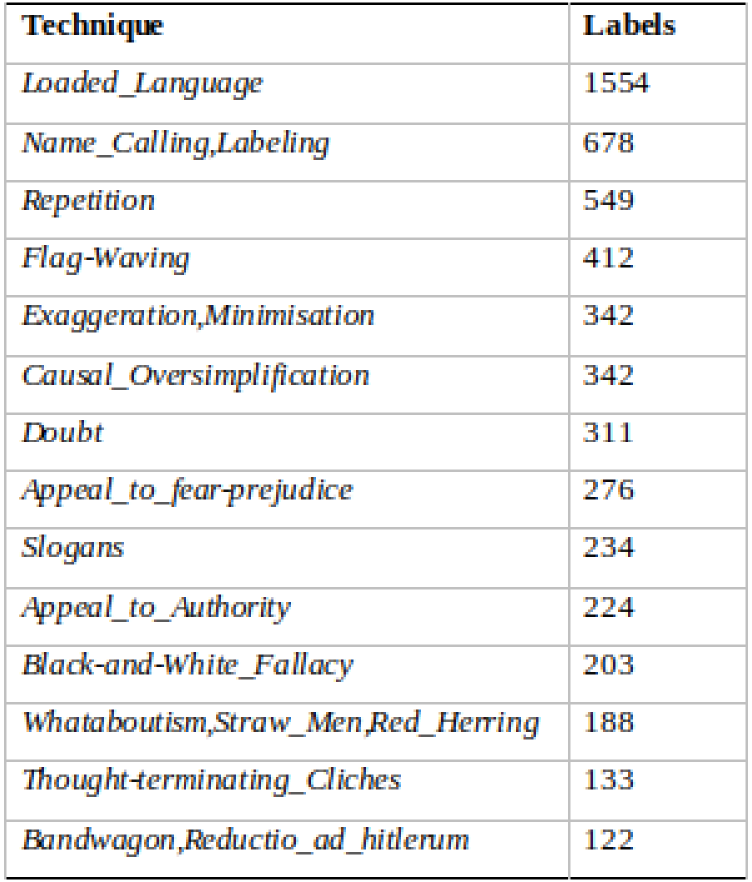
Span Identification(SI): Given a plain-text document, identify those specific fragments which are propagandistic.
Technique Classification(TC): Given a text fragment identified as propaganda and its document context, identify the applied propaganda technique in the fragment.

Dataset

- Number of articles: 550;
- Number of sentences: 16,297;
- Sentences labeled as “propaganda”: 4,720;
- Sentences labeled as “non-propaganda”: 11,577.

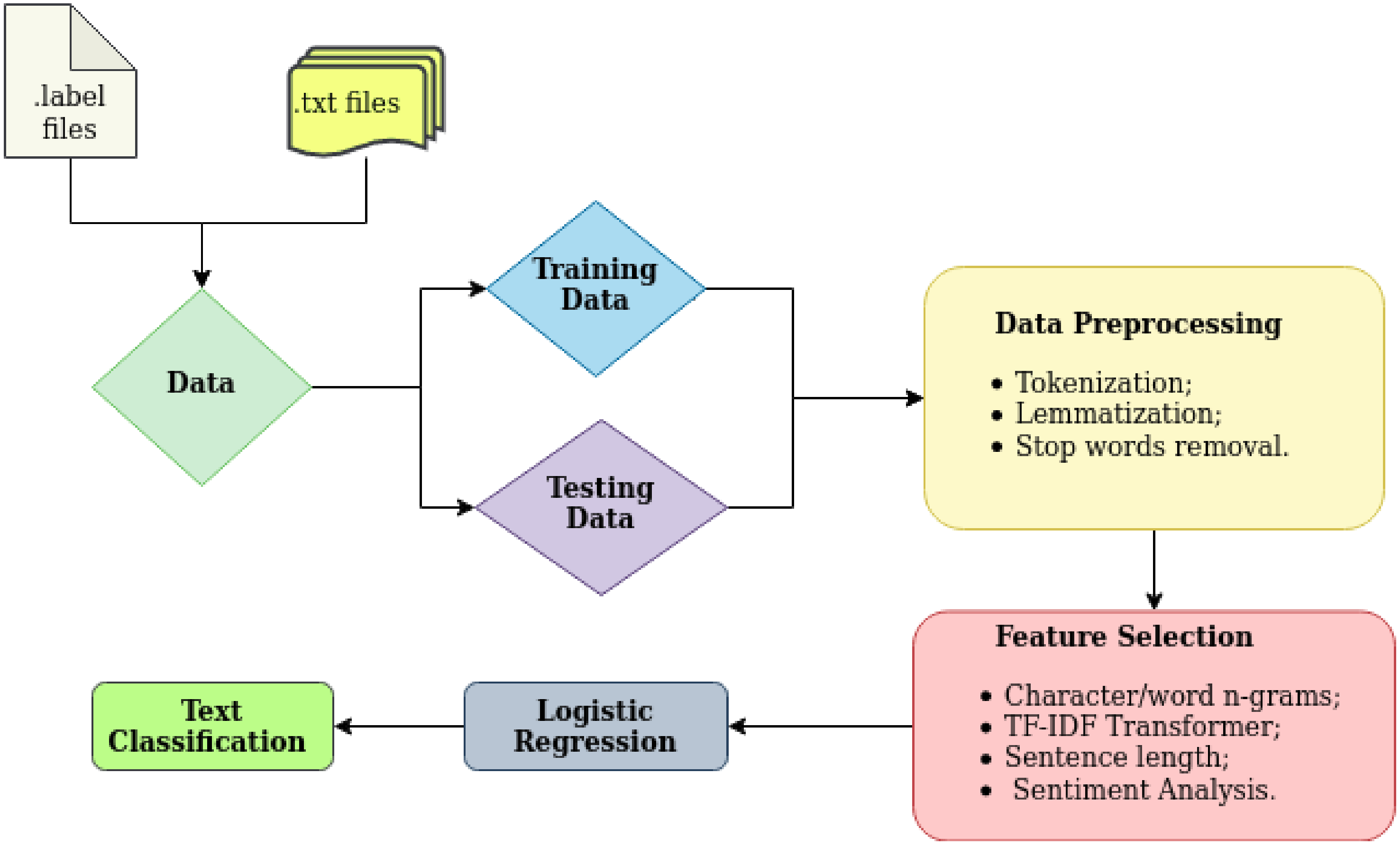
Distribution of the classes





Technique	**Labels**
Loaded_Language	1554
Name_Calling,Labeling	678
Repetition	549
Flag-Waving	412
Exaggeration,Minimisation	342
Causal_Oversimplification	342
Doubt	311
Appeal_to_fear-prejudice	276
Slogans	234
Appeal_to_Authority	224
Black-and-White_Fallacy	203
Whataboutism,Straw_Men,Red_Herring	188
Thought-terminating_Cliches	133
Bandwagon,Reductio_ad_hitlerum	122

System architecture. A proposal



Data Preprocessing

- Tokenization;
- Lemmatization;
- Stop words removal.

Feature Selection

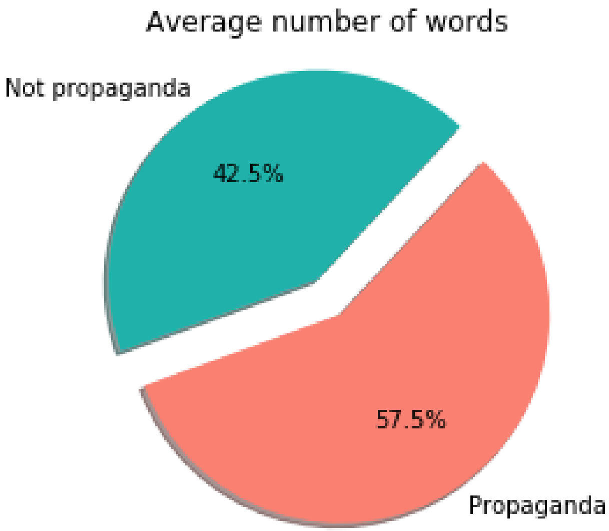
- Character/word n-grams;
- TF-IDF Transformer;
- Sentence length;
- Sentiment Analysis.

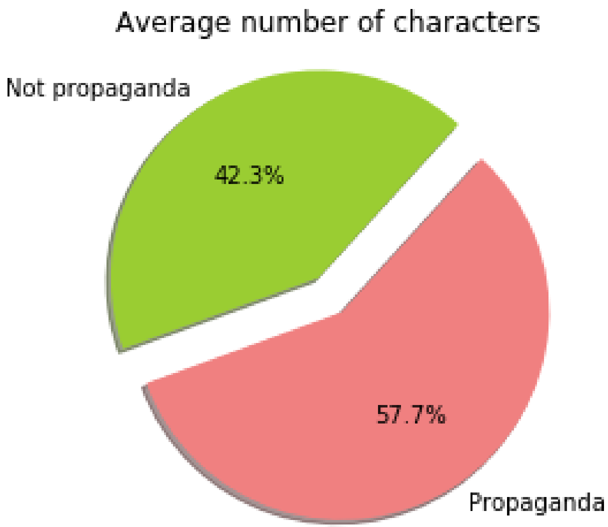
Dataset Preprocessing

- Tokenization;
- Lowercasing;
- Lemmatization;
- Removing special characters;
- Removing stop words.

Data Analysis

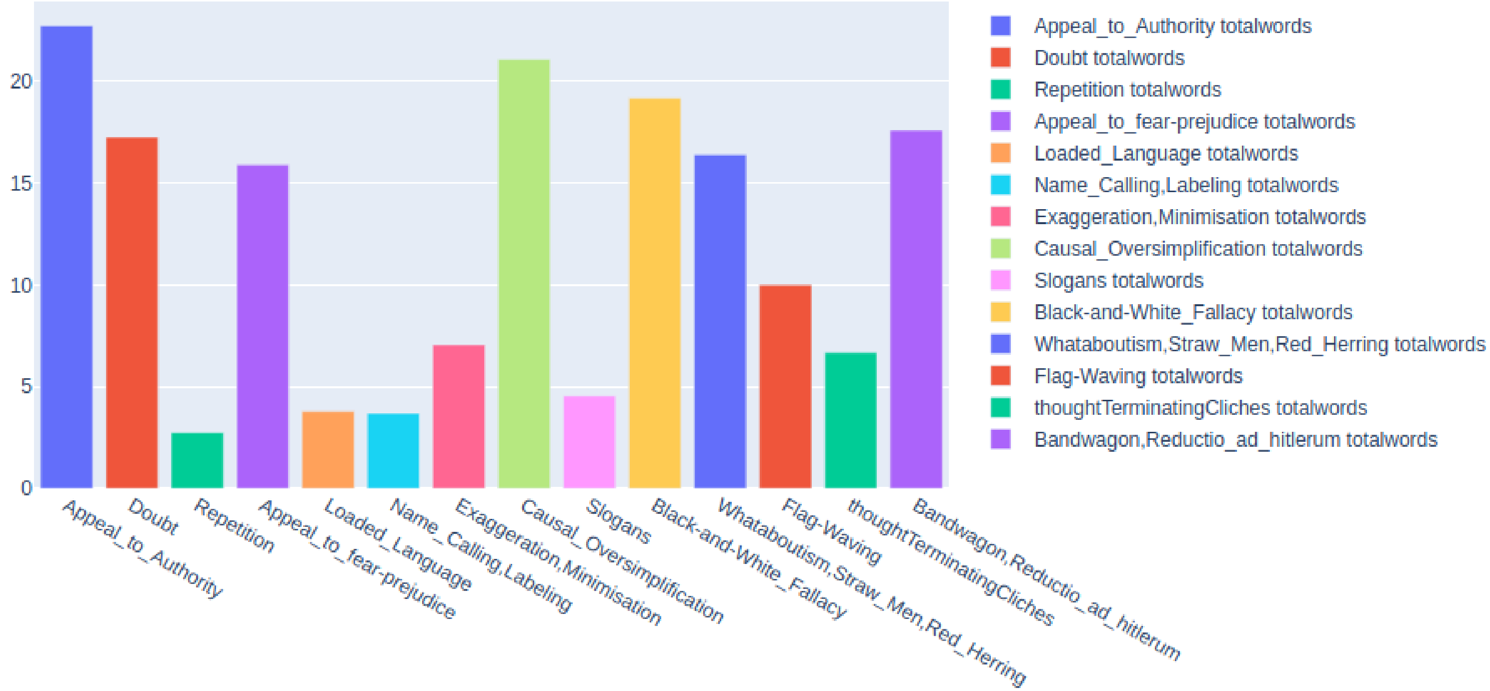
The average length of sentences containing propaganda is 25.16 words or 152.17 characters, while for sentences without propaganda this values are 18.61 and 112.06 respectively.





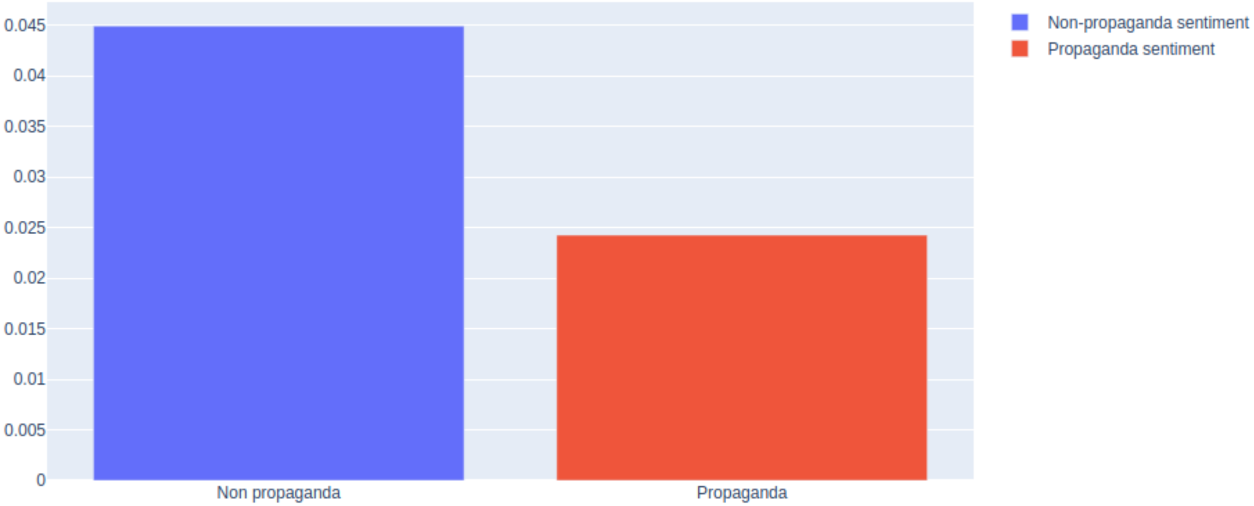
Data Analysis

Taking a closer look to the sentence length in different classes, we can recognize the techniques that implies long phrases to confuse the reader (e.g. *Appeal_to_authority*, *Casual_oversimplification*) or very short that intends to be catchy and memorable (e.g. *Repetition*, *Slogans*, *Loaded_Language*).



Polarity Measures

Performing a polarity measure of the sentence where 1 is positive and -1 is negative, it became obvious that negative sentiments are prevailed in the phrases containing propaganda, while *non-propaganda* instances tend to be loaded with neutral to positive sentiments.

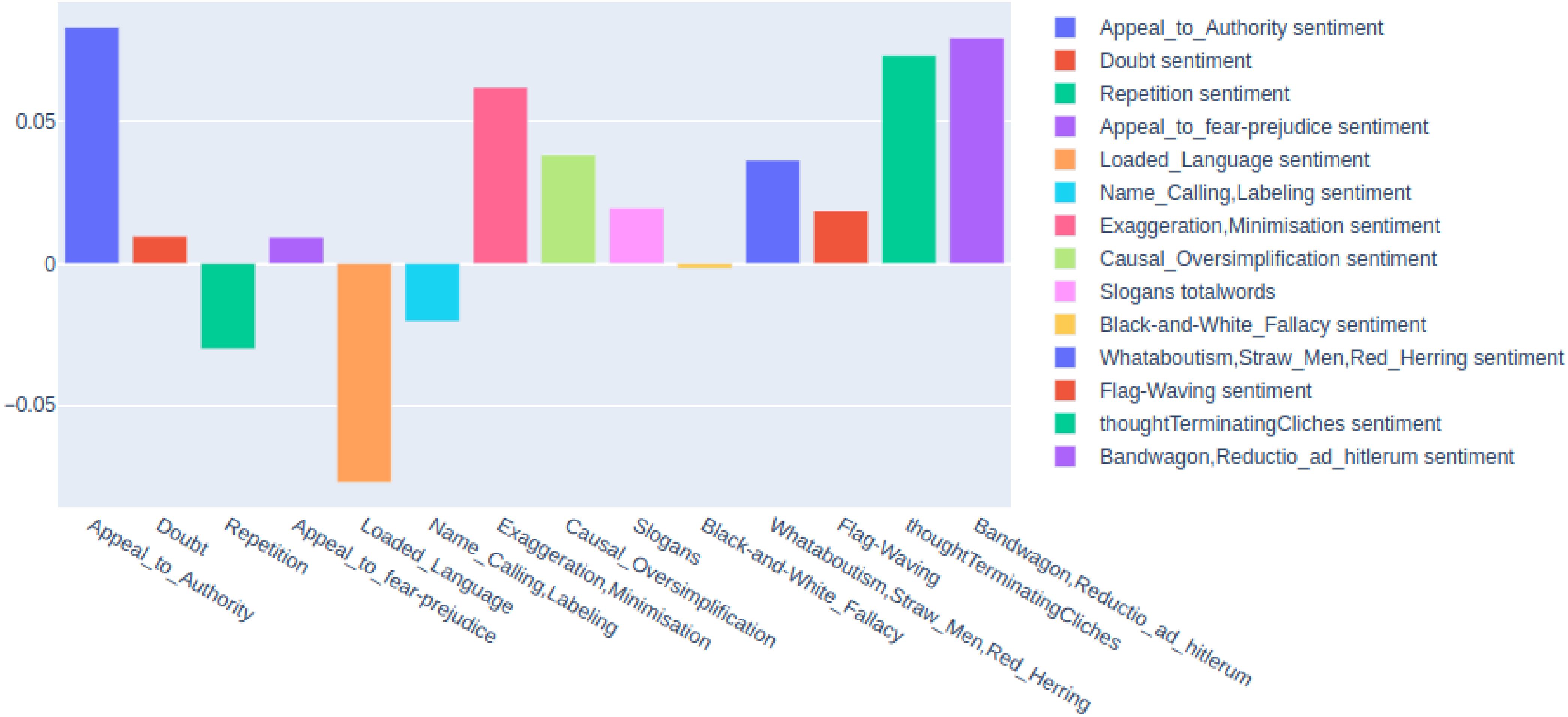


Polarity Measure for Each Class

Most of the techniques don’t illustrate an evident presence of negativity.

Ex: *Appeal_to_Authority*, *Exaggeration*,*Minimisation*, *Thought-terminating_Cliches*), however this does not mean that they are positive either, since their messages can be embedded to go unnoticed. Simultaneously, techniques like *Loaded_Language* and *Name_Calling*,*Labelig* are obviously charged with negative words that offsets their polarity scores.

Polarity Measure for Each Class

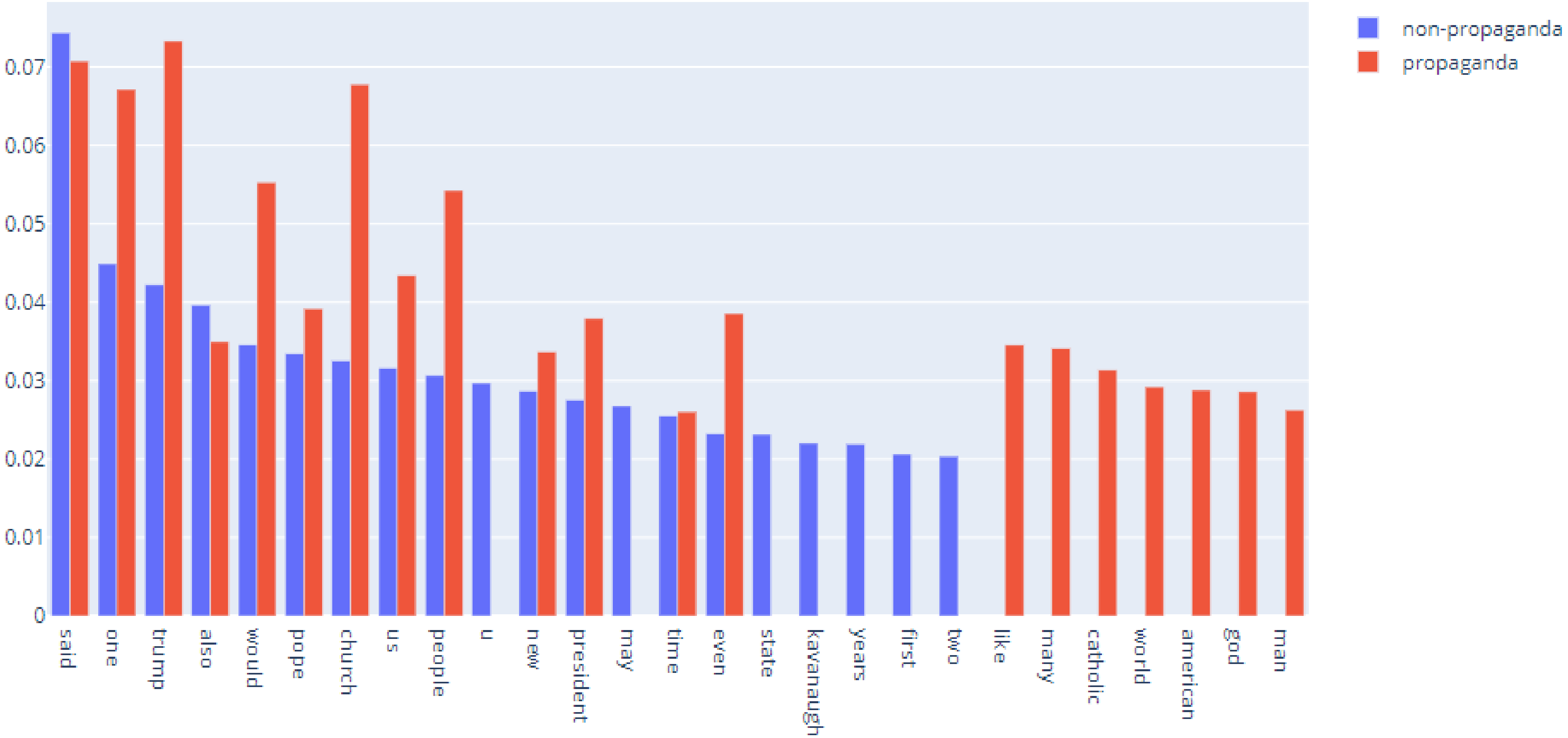


Most Frequent Words

Plotting the most common terms of the opposite classes in the same graphic we can observe their distribution in comparison.

Words like: *Trump*, *church*, *people* are almost twice as frequent in *propaganda* sentences as in *non-propaganda*. Another point to be made is that some words are almost absent in the opposite class: *world*, *American*, *god*, *man* are all characteristic for phrases in which propaganda was spotted.

Most Frequent Words



Features

- Bag of Words [1-4] character and [1-3] word n-grams;
- Term Frequency-Inverse Document Frequency;
- Part of speech tags;
- Named entity recognition;
- Numeric features (sent. length, sentiment).

Bag-of-Words

Bag-of-Words (n-grams) - one of the most frequent techniques of text representations used in many NLP tasks (e.g. IR; Q/A), where each dimension represents the number of occurrences of a word in a text (Bofang et al. 2016).

Could be done several experiments with different word and character n-grams, starting with unigrams and increasing their range up to 6-grams.

Term Frequency – Inverse Document Frequency

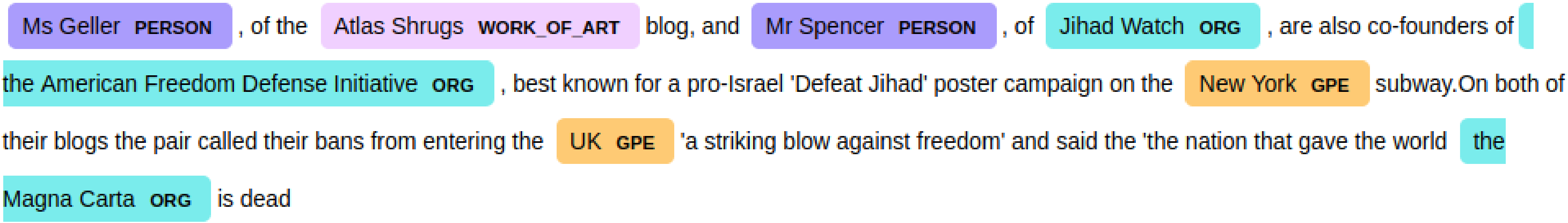
The reason of using **tf*idf** instead of the raw number of occurrence of a token in a given text is to scale down the influence of tokens that appear very often in the provided corpus and thus are generally less informative than features that occur in a smaller fragment of the training corpus.

Part of speech tags

POS-tags can provide valuable semantic information and outline specific syntactic patterns in sentences. We applied a pre-trained English pos-tagger provided by *spcay* to count occurrences of different parts of speech in our corpus. This aided us to determine the number of nouns, verbs and adjectives in different fragments, and use these numbers as features.

Named Entity Recognition

As our analysis shows that many biased sentences contain names of public persons and organizations, we decided that these feature might be valuable in identifying propaganda.



Ms Geller PERSON , of the Atlas Shrugs WORK_OF_ART blog, and Mr Spencer PERSON , of Jihad Watch ORG , are also co-founders of the American Freedom Defense Initiative ORG , best known for a pro-Israel 'Defeat Jihad' poster campaign on the New York GPE subway.On both of their blogs the pair called their bans from entering the UK GPE 'a striking blow against freedom' and said the 'the nation that gave the world the Magna Carta ORG is dead

Numeric features

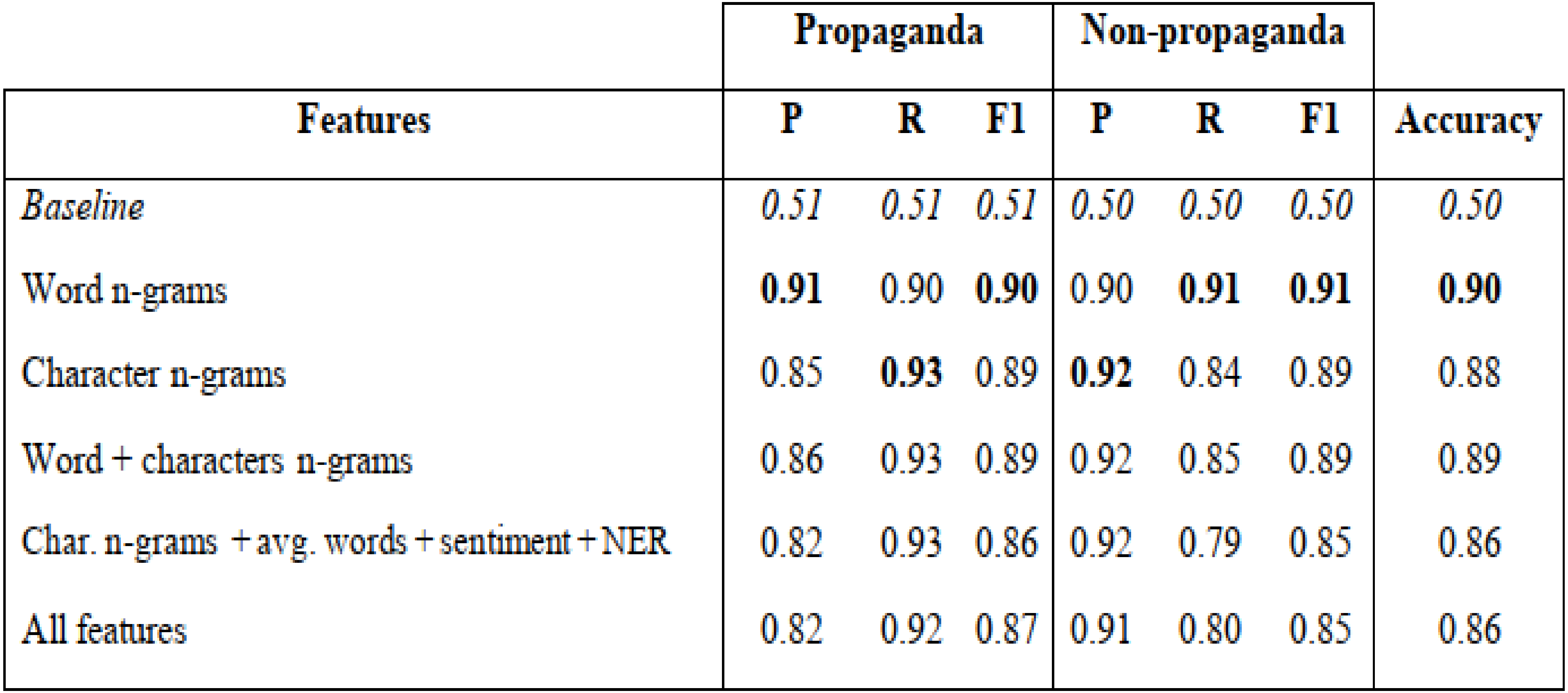
Several experiments were performed using:

- sentences length in words/characters;
- average word length in sentences;
- polarity scores;
- number of upper case words;
- number of punctuation marks.

Classification Algorithm

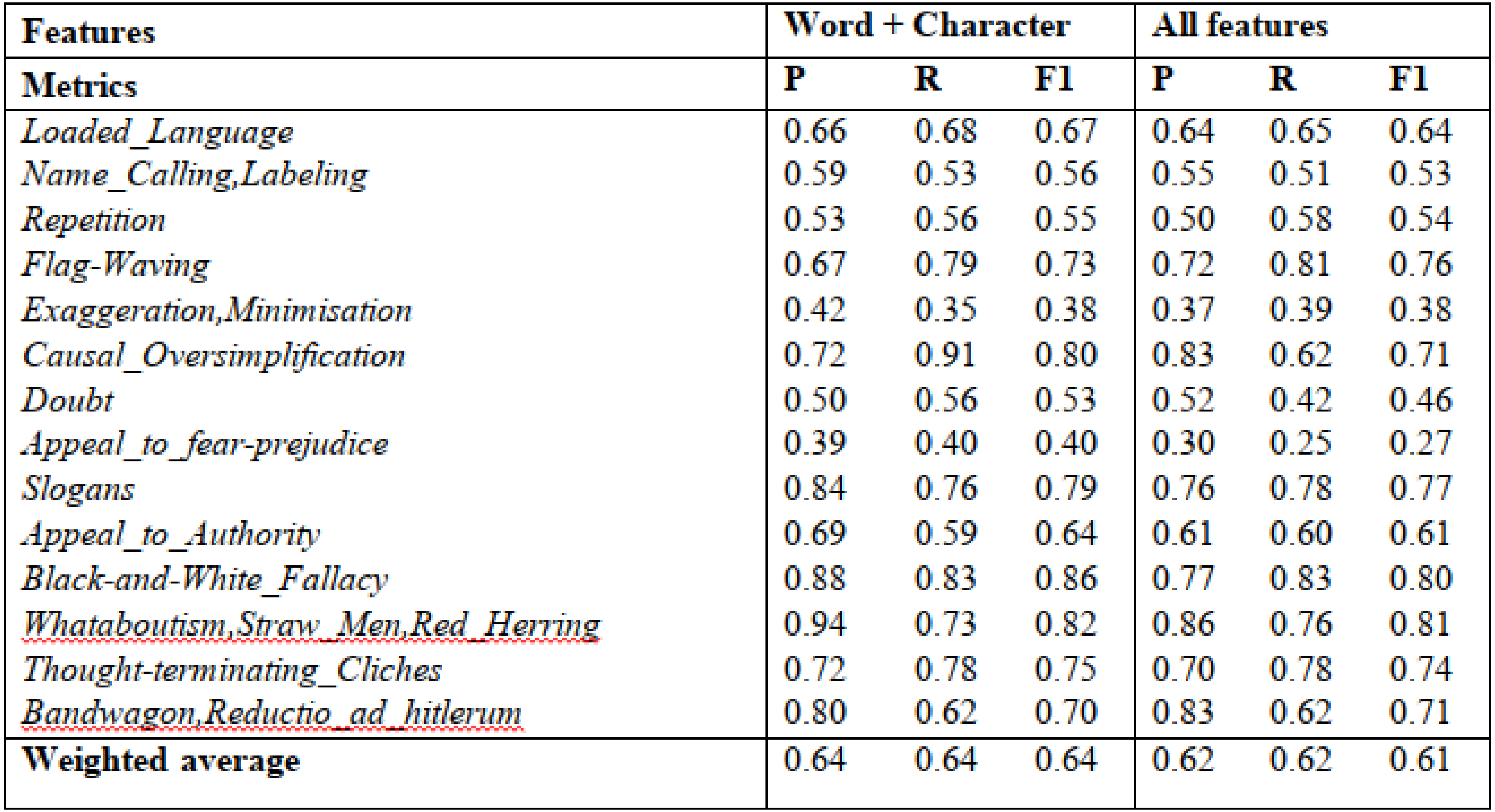
We have mainly focused on the Logistic Regression classifier, as it is one of the most frequently used supervised machine learning algorithm for classification, in natural language processing. Mostly it is applied on binary classification problems (e.g. ‘positive’/ ‘negative’ sentiment, or ‘propaganda’/‘non-propaganda’ as in our case), but can be used for multinational classifications as well, applying one-vs-all technique.

Interpretation of statistical results



Features	Propaganda			Non-propaganda			
	P	**R**	**F1**	**P**	**R**	**F1**	**Accuracy**
Baseline	*0.51*	*0.51*	*0.51*	*0.50*	*0.50*	*0.50*	*0.50*
Word n-grams	**0.91**	0.90	**0.90**	0.90	**0.91**	**0.91**	**0.90**
Character n-grams	0.85	**0.93**	0.89	**0.92**	0.84	0.89	0.88
Word + characters n-grams	0.86	0.93	0.89	0.92	0.85	0.89	0.89
Char. n-grams +avg. words + sentiment + NER	0.82	0.93	0.86	0.92	0.79	0.85	0.86
All features	0.82	0.92	0.87	0.91	0.80	0.85	0.86

Interpretation of statistical results



Features	**Word + Character**			**All features**		
Metrics	**P**	**R**	**F1**	**P**	**R**	**F1**
Loaded_Language	0.66	0.68	0.67	0.64	0.65	0.64
Name_Calling,Labeling	0.59	0.53	0.56	0.55	0.51	0.53
Repetition	0.53	0.56	0.55	0.50	0.58	0.54
Flag-Waving	0.67	0.79	0.73	0.72	0.81	0.76
Exaggeration,Minimisation	0.42	0.35	0.38	0.37	0.39	0.38
Causal_Oversimplification	0.72	0.91	0.80	0.83	0.62	0.71
Doubt	0.50	0.56	0.53	0.52	0.42	0.46
Appeal_to_fear-prejudice	0.39	0.40	0.40	0.30	0.25	0.27
Slogans	0.84	0.76	0.79	0.76	0.78	0.77
Appeal_to_Authority	0.69	0.59	0.64	0.61	0.60	0.61
Black-and-White_Fallacy	0.88	0.83	0.86	0.77	0.83	0.80
Whataboutism,Straw_Men,Red_Herring	0.94	0.73	0.82	0.86	0.76	0.81
Thought-terminating_Cliches	0.72	0.78	0.75	0.70	0.78	0.74
Bandwagon,Reductio_ad_hitlerum	0.80	0.62	0.70	0.83	0.62	0.71
Weighted average	0.64	0.64	0.64	0.62	0.62	0.61

Conclusion I

- nature of propaganda with the implication of NLP/AI techniques.

- build a model capable of identifying and classifying instances of propaganda.

- build the application interface for a more convenient use.

Conclusion II

- Need a larger and better annotated corpus that would provide more opportunities for exploring the issue;
- Training and testing DL.

See you next class!