Optimising Agile Social Media Analysis

Thomas Kober
t.kober@sussex.ac.uk

David Weir
d.j.weir@sussex.ac.uk

University of Sussex

Outline

- Introduction & Methodology
- Practical Aspects
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

Outline

- ➡ **Introduction & Methodology**
- Practical Aspects
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

Introduction

- Agile Social Media Analysis
  - Building *bespoke* classifiers for performing *specific* analyses on *user-defined* topics on large social media datasets.

Introduction

- Agile Social Media Analysis
    - Building *bespoke* classifiers for performing *specific* analyses on *user-defined* topics on large social media datasets.
- Probably better explained with an example…

Agile Social Media Analysis

Agile Social Media Analysis

- A typical scenario…
  - …involves a “Twitcident”, e.g. a political leader giving a speech

Agile Social Media Analysis

- A typical scenario…
  - …involves a "Twitcident", e.g. a political leader giving a speech
- The goal is to analyse the reactions to this speech
  - What contents caused the most controversy?
  - Why are these topics so fiercely debated?
  - Are reactions to a specific topic mostly positive or negative?

Agile Social Media Analysis

Agile Social Media Analysis

- A political scientist wants to analyse the reactions to a speech given by British Prime Minister David Cameron the previous night

Agile Social Media Analysis

- A political scientist wants to analyse the reactions to a speech given by British Prime Minister David Cameron the previous night
- She queries the Twitter API with “Cameron” to retrieve an initial dataset

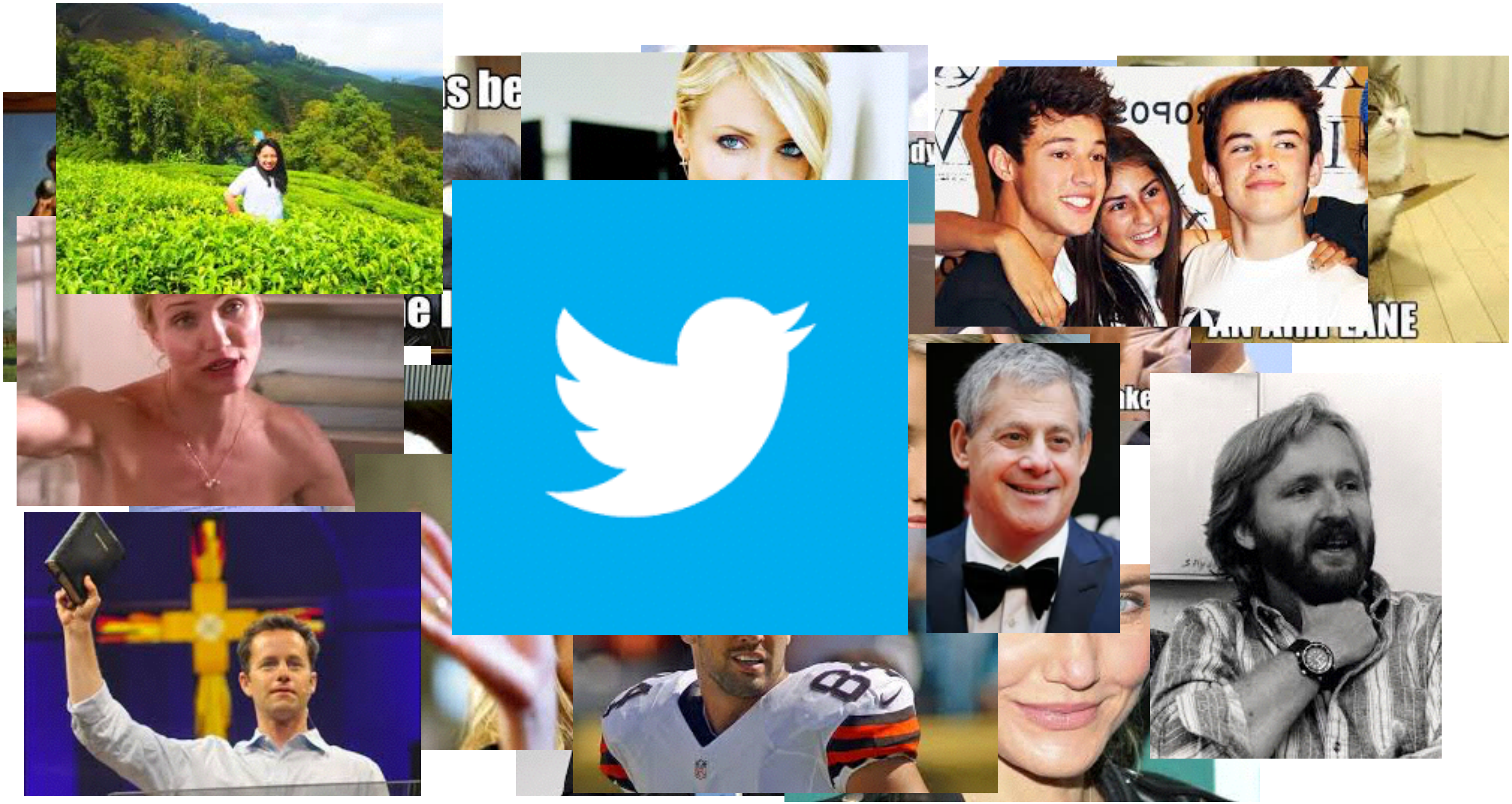
Agile Social Media Analysis

Agile Social Media Analysis

Agile Social Media Analysis

Agile Social Media Analysis

Agile Social Media Analysis



Agile Social Media Analysis

Agile Social Media Analysis

- In the beginning the dataset is a “heterogenous mass of text”

Agile Social Media Analysis

- In the beginning the dataset is a “heterogenous mass of text”
- Very limited appreciation of the contents of the data in the beginning

Agile Social Media Analysis

- In the beginning the dataset is a “heterogenous mass of text”
- Very limited appreciation of the contents of the data in the beginning
- No labelled data

Agile Social Media Analysis

- In the beginning the dataset is a “heterogenous mass of text”
- Very limited appreciation of the contents of the data in the beginning
- No labelled data
- No off the shelf dataset/classifier that can be used for the target analysis

Supervised Machine Learning meets agile Social Media Analysis…

Agile Social Media Analysis

Agile Social Media Analysis

- Performing Sentiment Analysis on this initially retrieved dataset will give poor results

Agile Social Media Analysis

- Performing Sentiment Analysis on this initially retrieved dataset will give poor results
- And more importantly, no actual insight into people's reaction to the debate

Agile Social Media Analysis

- Performing Sentiment Analysis on this initially retrieved dataset will give poor results
- And more importantly, no actual insight into people's reaction to the debate
- Need a tailored multi-stage processing pipeline and direct interaction with the data

Agile Social Media Analysis

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
    - This classifier will only be used for *this single task*

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
    - This classifier will only be used for *this single task*
    - Start with a random sample from the initially retrieved tweets and label a gold standard set

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
    - This classifier will only be used for *this single task*
    - Start with a random sample from the initially retrieved tweets and label a gold standard set
    - Labelling a gold standard set also serves to *explore the data space*

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
  - This classifier will only be used for *this single task*
  - Start with a random sample from the initially retrieved tweets and label a gold standard set
  - Labelling a gold standard set also serves to *explore the data space*
  - The prevalent subtopics “personality”, “migrant crisis” and “EU referendum” are identified

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
    - This classifier will only be used for *this single task*
    - Start with a random sample from the initially retrieved tweets and label a gold standard set
    - Labelling a gold standard set also serves to *explore the data space*
    - The prevalent subtopics “personality”, “migrant crisis” and “EU referendum” are identified
    - After the gold standard is labelled, active learning is used to train a classifier

Agile Social Media Analysis

- First step is to train a classifier for relevancy classification
    - This classifier will only be used for *this single task*
    - Start with a random sample from the initially retrieved tweets and label a gold standard set
    - Labelling a gold standard set also serves to *explore the data space*
    - The prevalent subtopics “personality”, “migrant crisis” and “EU referendum” are identified
    - After the gold standard is labelled, active learning is used to train a classifier
    - The classifier is applied to the dataset, only the relevant tweets are used for further processing steps

Agile Social Media Analysis

Agile Social Media Analysis

- Create another bespoke classifier to split the relevant tweets into the 3 identified prevalent subtopics

Agile Social Media Analysis

- Create another bespoke classifier to split the relevant tweets into the 3 identified prevalent subtopics
    - Same workflow, label a gold standard from a random sample, then use active learning to train a classifier

Agile Social Media Analysis

- Create another bespoke classifier to split the relevant tweets into the 3 identified prevalent subtopics
    - Same workflow, label a gold standard from a random sample, then use active learning to train a classifier
- Finally, Sentiment Analysis can be performed on each of the 3 subtopics separately

Agile Social Media Analysis

Agile Social Media Analysis

- The result is a highly specialised classification pipeline tailored for a specific and granular analysis of an event

Agile Social Media Analysis

- The result is a highly specialised classification pipeline tailored for a specific and granular analysis of an event
- Direct Interaction with the data is crucial
    - Discover what the data is about
    - Tailor the analysis to the given data

Agile Social Media Analysis

- The result is a highly specialised classification pipeline tailored for a specific and granular analysis of an event
- Direct Interaction with the data is crucial
    - Discover what the data is about
    - Tailor the analysis to the given data
- Fast hypothesis testing
    - System reports performance on gold standard set after each retraining step
    - “Fail Fast” if the data doesn’t align with the target labels

Outline

- Introduction & Methodology
- Practical Aspects
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

Outline

- Introduction & Methodology
- ➡ **Practical Aspects**
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

The System

The System

- General classifier training architecture based on DUALIST
    - Combines a classifier, a semi-supervised learning algorithm and active learning into an application

The System

- General classifier training architecture based on DUALIST
    - Combines a classifier, a semi-supervised learning algorithm and active learning into an application
- Our system, **method51**, has been extended in several ways (Wibberley et al. 2013; Wibberley et al. 2014)
    - Querying the Twitter API
    - Gold Standard Sampling
    - Measuring Inter-Annotator Agreement
    - Classifier pipelining

The System

- General classifier training architecture based on DUALIST
    - Combines a classifier, a semi-supervised learning algorithm and active learning into an application
- Our system, **method51**, has been extended in several ways (Wibberley et al. 2013; Wibberley et al. 2014)
    - Querying the Twitter API
    - Gold Standard Sampling
    - Measuring Inter-Annotator Agreement
    - Classifier pipelining
- New bespoke classifiers can be built in ~15-30mins

method51 - Classifier Pipeline

method51 - Classifier Pipeline

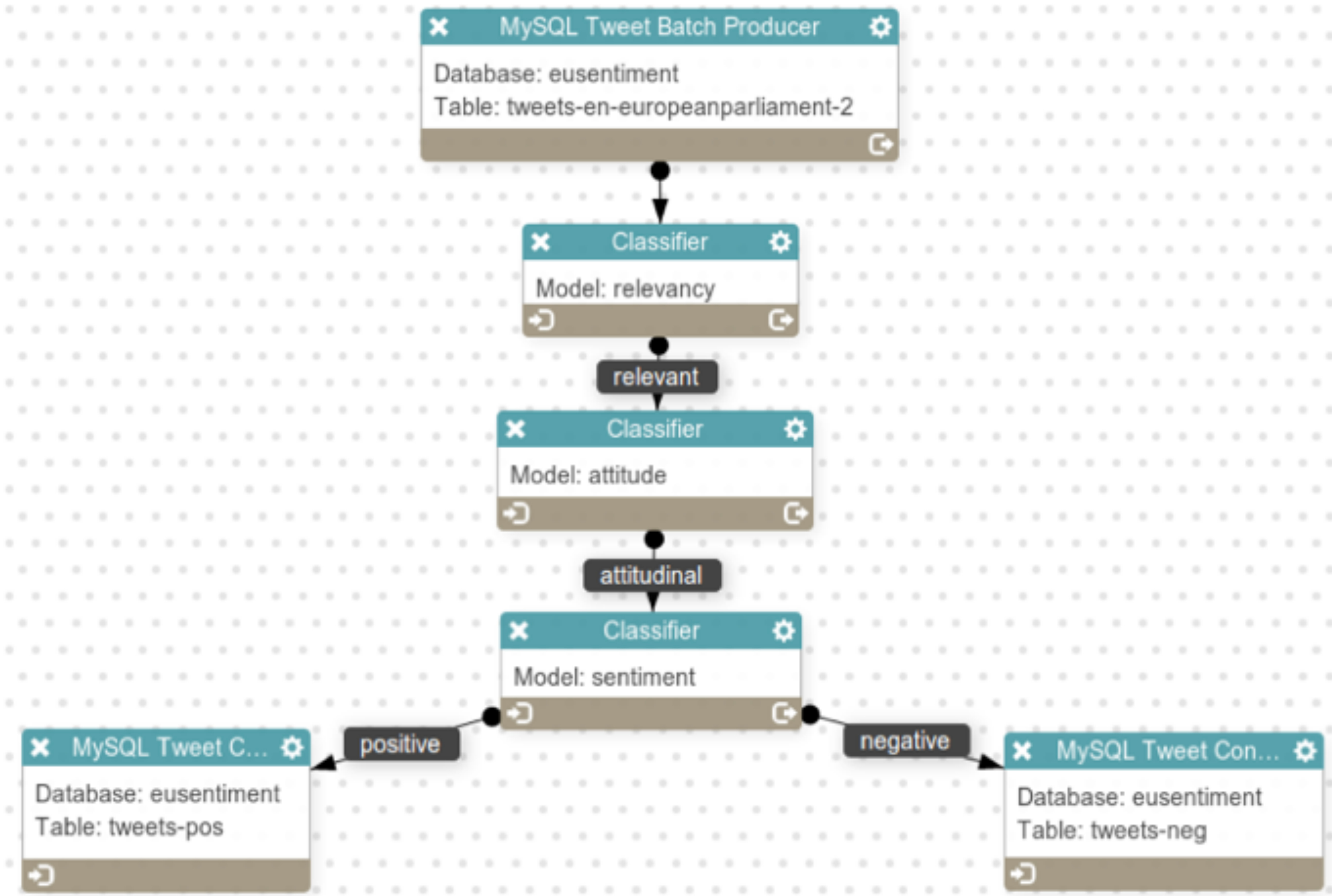


Figure 1: Processing Pipeline Interface

method51 - Classifier Training

Label		Precision	Recall	F-Score	Accuracy	Coded	Label Multiplier	Alpha	Action
positive	Sample	0.447	0.602	0.513		724	1		Process
negative	Sample	0.768	0.849	0.806		1371	1		Process
neutral	Sample	0.667	0.305	0.419		519	1		Process
Unlabelled		9716	**Features**	76	0.672		Standard EM	10	sent out 123

10 records per page

Search:

Showing 611 to 620 of 2,624 entries

← Previous 60 61 62 63 64 Next →

Document	
And farage responds with Latin. Nice #LBCdebate	positive negative neutral
@Nigel_Farage go for it, most of us (workers) we want out.	positive negative neutral
Did Farage have a curry before he went on? Sweating buckets. #LBCdebate	positive negative neutral
Farage already losing this debate. Speaking too fast and not engaging with the camera #LBCdebate	positive negative neutral

Figure 2: Classifier Training Interface

method51 - Classifier Training

method51 - Classifier Training

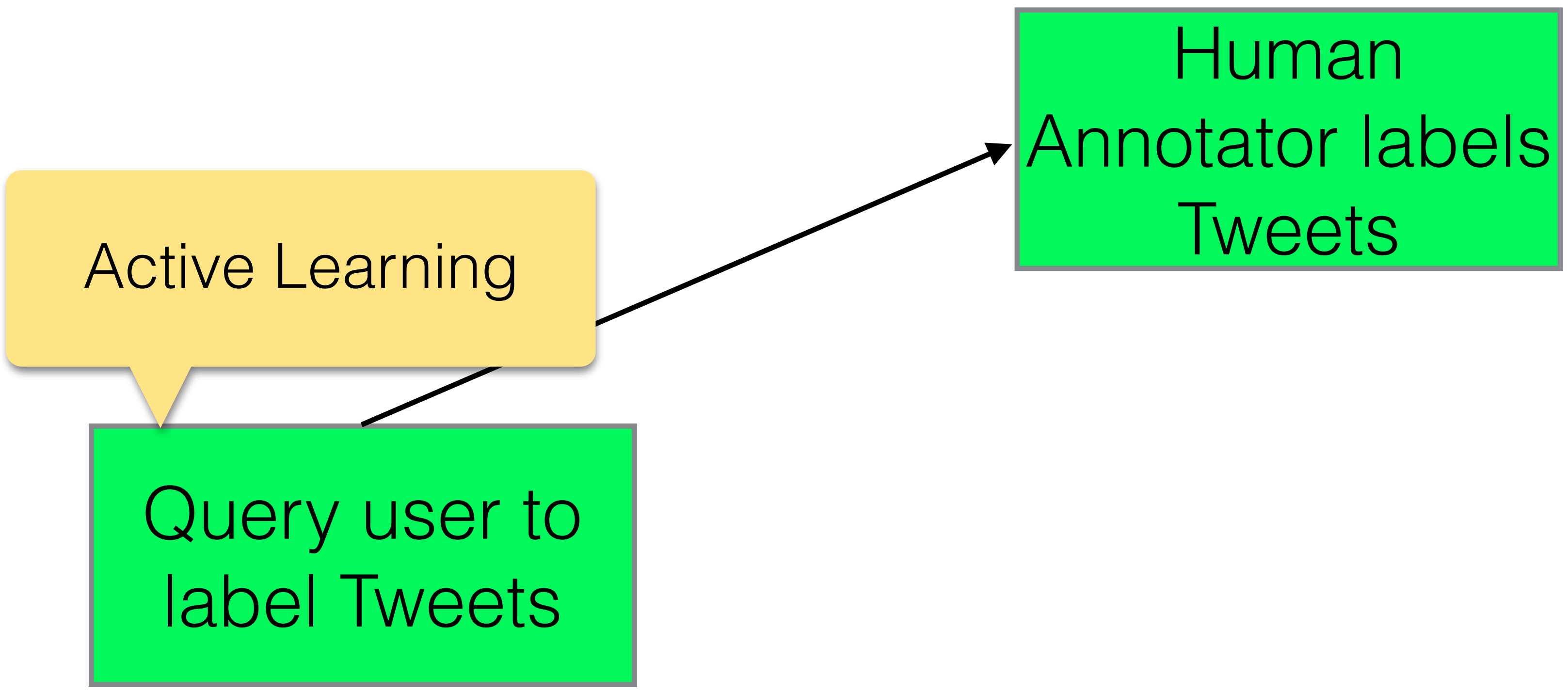
Query user to label Tweets

method51 - Classifier Training

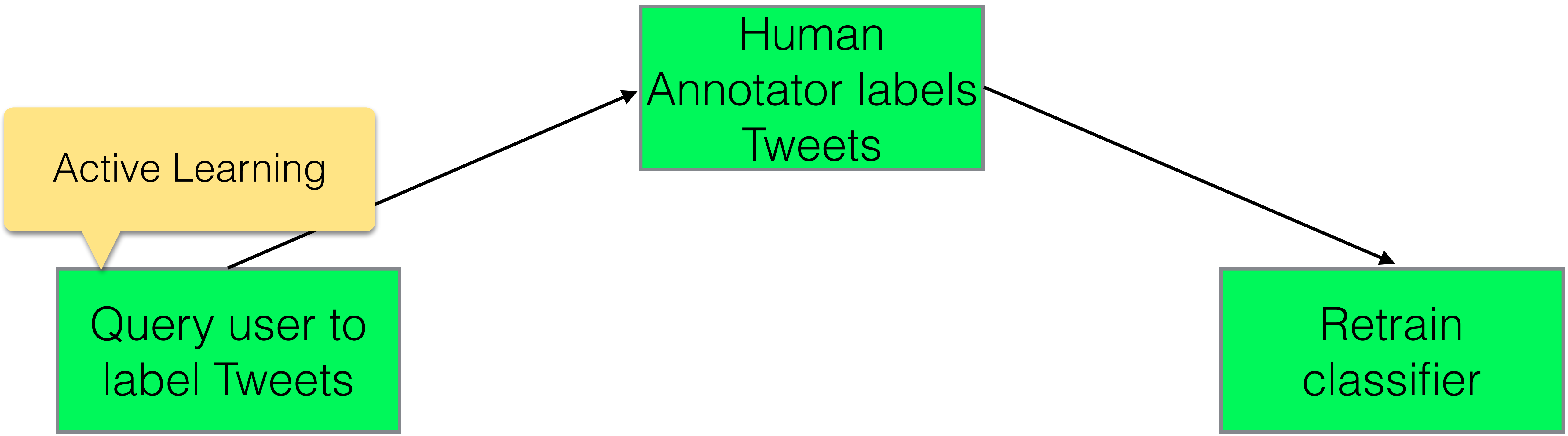
Active Learning

Query user to label Tweets

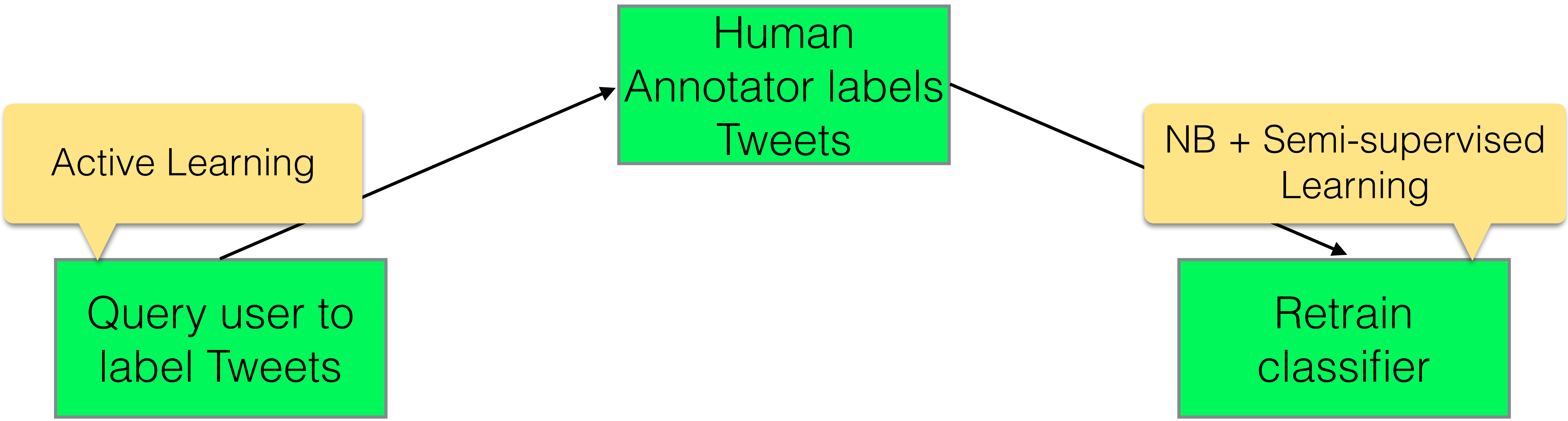
method51 - Classifier Training



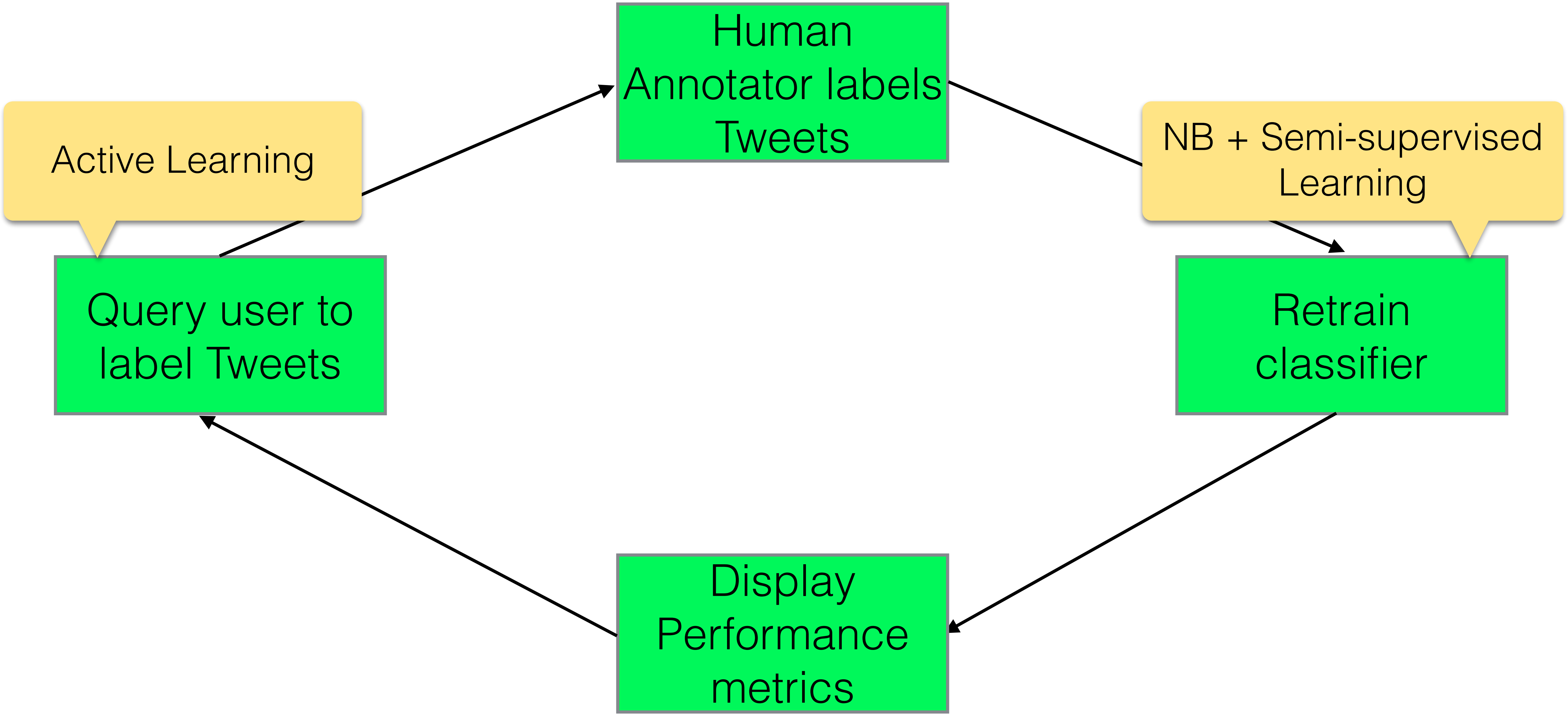
method51 - Classifier Training



method51 - Classifier Training



method51 - Classifier Training



Outline

- Introduction & Methodology
- Practical Aspects
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

Outline

- Introduction & Methodology
- Practical Aspects

➡ **Optimising Agile Social Media Analysis**

- Conclusion & Outlook

Focus of this Paper

Focus of this Paper

- Optimise the classification engine (classifier and semi-supervised learning algorithm)

Focus of this Paper

- Optimise the classification engine (classifier and semi-supervised learning algorithm)

- *Major challenge:* Improve classification effectiveness by maintaining real-time user-interaction

Experimental Setup

Experimental Setup

- Baseline: Multinomial NB + Expectation Maximization

Experimental Setup

- Baseline: Multinomial NB + Expectation Maximization
- Parameterisation and selection of the Naive Bayes event model
    - Bernoulli
    - Multinomial
    - binary Multinomial

Experimental Setup

- Baseline: Multinomial NB + Expectation Maximization
- Parameterisation and selection of the Naive Bayes event model
  - Bernoulli
  - Multinomial
  - binary Multinomial
- Semi-supervised learning algorithms comparison
  - Expectation-Maximization (EM); e.g. Nigam et al. (1999)
  - Semi-supervised Frequency Estimate (SFE); Su et al. (2011)
  - Feature Marginals (FM); Lucas & Downey (2013)

Experimental Setup

Experimental Setup

- The effect of unlabelled data
    - Do unlabelled data help?
    - How much of the unlabelled data is necessary?

Experimental Setup

- The effect of unlabelled data
  - Do unlabelled data help?
  - How much of the unlabelled data is necessary?
- The effect of adding bigrams and trigrams

Evaluation

Evaluation

- 24 Twitter Datasets
  - 12 Topic Classification, 12 Sentiment Analysis
  - Variety of topics ranging from political debates and extremism to natural disasters (among others)
  - Very few labelled data (~hundreds)
  - Large amount of unlabelled data (~tens to hundreds of thousands)

Evaluation

- 24 Twitter Datasets
  - 12 Topic Classification, 12 Sentiment Analysis
  - Variety of topics ranging from political debates and extremism to natural disasters (among others)
  - Very few labelled data (~hundreds)
  - Large amount of unlabelled data (~tens to hundreds of thousands)
- Movie Reviews (Maas et al. 2011)

Evaluation

- 24 Twitter Datasets
  - 12 Topic Classification, 12 Sentiment Analysis
  - Variety of topics ranging from political debates and extremism to natural disasters (among others)
  - Very few labelled data (~hundreds)
  - Large amount of unlabelled data (~tens to hundreds of thousands)
- Movie Reviews (Maas et al. 2011)
- 20 Newsgroups (Lang 1995)

Findings & Results

Findings & Results

- Parameterisation and selection of the Naive Bayes event model
  - binary Multinomial NB and Bernoulli NB typically outperform Multinomial NB
  - No clear winner between binary Multinomial NB and Bernoulli NB
  - Findings align with previous studies

Findings & Results

- Parameterisation and selection of the Naive Bayes event model
    - binary Multinomial NB and Bernoulli NB typically outperform Multinomial NB
    - No clear winner between binary Multinomial NB and Bernoulli NB
    - Findings align with previous studies
- Semi-supervised learning algorithms comparison
    - SFE and FM outperform our EM baseline
    - EM with weighting heuristic is competitive with (and often superior to) SFE and FM
    - Baseline configuration outperformed on 24 out of 26 datasets (Performance gains up to 25%)

Findings & Results

Findings & Results

- Semi-supervised learning algorithms comparison

Findings & Results

- Semi-supervised learning algorithms comparison
  - Bad performance of Baseline EM configuration mainly due to assigning too much weight to the unlabelled data

Findings & Results

- Semi-supervised learning algorithms comparison
    - Bad performance of Baseline EM configuration mainly due to assigning too much weight to the unlabelled data
    - Performance among other algorithms inconsistent (differences of up to ~8% between algorithms on the same dataset)

Findings & Results

- Semi-supervised learning algorithms comparison
  - Bad performance of Baseline EM configuration mainly due to assigning too much weight to the unlabelled data
  - Performance among other algorithms inconsistent (differences of up to ~8% between algorithms on the same dataset)
  - Not entirely clear if it is a data or a hyper-parameter phenomenon

Findings & Results

Findings & Results

- The effect of unlabelled data

Findings & Results

- The effect of unlabelled data
    - Adding unlabelled data typically improves performance over a purely supervised approach (but not always!)

Findings & Results

- The effect of unlabelled data
  - Adding unlabelled data typically improves performance over a purely supervised approach (but not always!)
  - The *amount* of unlabelled data being added can have a significant effect

Findings & Results

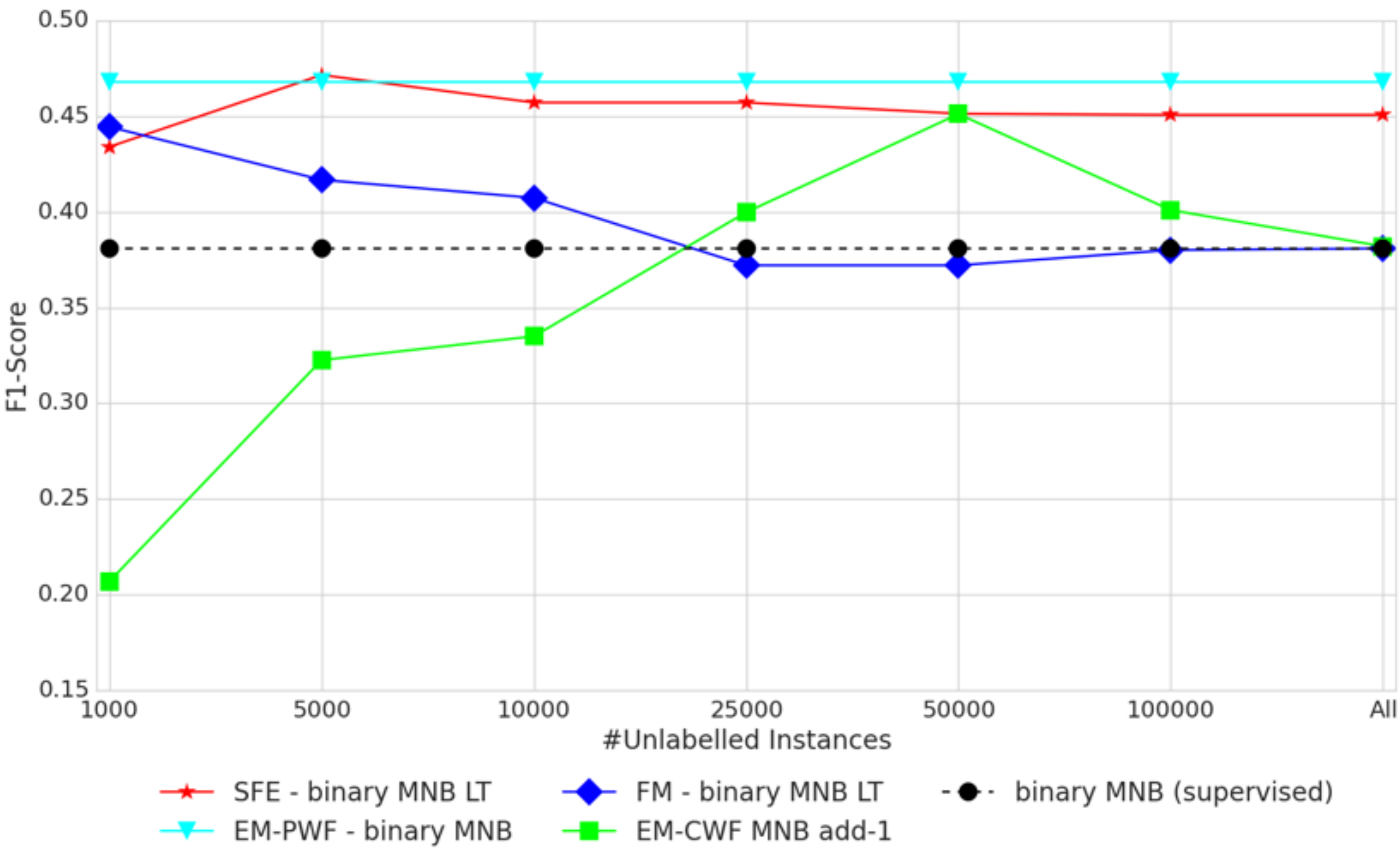


Figure 4: The effect of unlabelled data

Findings & Results

Findings & Results

- The effect of unlabelled data

Findings & Results

- The effect of unlabelled data
  - Performance of baseline configuration sensitive to amount of data

Findings & Results

- The effect of unlabelled data
    - Performance of baseline configuration sensitive to amount of data
    - SFE & FM stabler, but also show sensitivity to amount of data

Findings & Results

- The effect of unlabelled data
  - Performance of baseline configuration sensitive to amount of data
  - SFE & FM stabler, but also show sensitivity to amount of data
  - EM with weighting heuristic very stable

Findings & Results

- The effect of unlabelled data
    - Performance of baseline configuration sensitive to amount of data
    - SFE & FM stabler, but also show sensitivity to amount of data
    - EM with weighting heuristic very stable
    - *Too* stable - unlabelled data not used effective enough

Findings & Results

Findings & Results

- The effect of adding bigrams and trigrams

Findings & Results

- The effect of adding bigrams and trigrams
    - Contrary to a recent study, we did not observe any consistent improvements by adding bigrams and trigrams in our datasets (neither for Topic Classification, nor for Sentiment Analysis)

Findings & Results

- The effect of adding bigrams and trigrams
    - Contrary to a recent study, we did not observe any consistent improvements by adding bigrams and trigrams in our datasets (neither for Topic Classification, nor for Sentiment Analysis)
    - We observed the inconsistent behaviour in both, supervised and semi-supervised settings

Findings & Results

- The effect of adding bigrams and trigrams
  - Contrary to a recent study, we did not observe any consistent improvements by adding bigrams and trigrams in our datasets (neither for Topic Classification, nor for Sentiment Analysis)
  - We observed the inconsistent behaviour in both, supervised and semi-supervised settings
  - A possible explanation could be the usage of multi-word hashtag expressions, e.g. “#CameronMustGo” or “#CareNotCuts”, which convey crucial sentiment information but are treated as unigrams

Findings & Results

- The effect of adding bigrams and trigrams
    - Contrary to a recent study, we did not observe any consistent improvements by adding bigrams and trigrams in our datasets (neither for Topic Classification, nor for Sentiment Analysis)
    - We observed the inconsistent behaviour in both, supervised and semi-supervised settings
    - A possible explanation could be the usage of multi-word hashtag expressions, e.g. “#CameronMustGo” or “#CareNotCuts”, which convey crucial sentiment information but are treated as unigrams
    - Similarly, the Topic Classification corpora also contained such multi-word expressions, e.g. “#ArcticOil”, that define the topic of a tweet

Findings & Results

- The effect of adding bigrams and trigrams
  - Contrary to a recent study, we did not observe any consistent improvements by adding bigrams and trigrams in our datasets (neither for Topic Classification, nor for Sentiment Analysis)
  - We observed the inconsistent behaviour in both, supervised and semi-supervised settings
  - A possible explanation could be the usage of multi-word hashtag expressions, e.g. "#CameronMustGo" or "#CareNotCuts", which convey crucial sentiment information but are treated as unigrams
  - Similarly, the Topic Classification corpora also contained such multi-word expressions, e.g. "#ArcticOil", that define the topic of a tweet
  - Therefore we hypothesise that bigrams and trigrams cannot be leveraged as effectively for Twitter datasets as for other datasets

Outline

- Introduction & Methodology
- Practical Aspects
- Optimising Agile Social Media Analysis
- Conclusion & Outlook

Outline

- Introduction & Methodology
- Practical Aspects
- Optimising Agile Social Media Analysis

➡ **Conclusion & Outlook**

Conclusion

Conclusion

- The aim of agile Social Media Analysis is to allow social scientists to build bespoke classifier pipelines to perform tailored analyses on large social media datasets

Conclusion

- The aim of agile Social Media Analysis is to allow social scientists to build bespoke classifier pipelines to perform tailored analyses on large social media datasets
- Our system leverages active learning and semi-supervised learning together with a Naive Bayes classifier to build custom classifiers for user-defined tasks

Conclusion

- The aim of agile Social Media Analysis is to allow social scientists to build bespoke classifier pipelines to perform tailored analyses on large social media datasets
- Our system leverages active learning and semi-supervised learning together with a Naive Bayes classifier to build custom classifiers for user-defined tasks
- With optimised hyper-parameters EM compares favourably to newer semi-supervised learning algorithms

Conclusion

- The aim of agile Social Media Analysis is to allow social scientists to build bespoke classifier pipelines to perform tailored analyses on large social media datasets
- Our system leverages active learning and semi-supervised learning together with a Naive Bayes classifier to build custom classifiers for user-defined tasks
- With optimised hyper-parameters EM compares favourably to newer semi-supervised learning algorithms
- Unlabelled data generally improve performance, but adding more data does not always mean better performance

Conclusion

- The aim of agile Social Media Analysis is to allow social scientists to build bespoke classifier pipelines to perform tailored analyses on large social media datasets
- Our system leverages active learning and semi-supervised learning together with a Naive Bayes classifier to build custom classifiers for user-defined tasks
- With optimised hyper-parameters EM compares favourably to newer semi-supervised learning algorithms
- Unlabelled data generally improve performance, but adding more data does not always mean better performance
- Bigrams and trigrams cannot be as effectively leveraged in Twitter datasets as in other datasets

Outlook

Outlook

- Investigation of the effect of different hyper-parameter settings
    - Can they be optimised automatically?
    - Can we find task or dataset invariant heuristics to optimise them?

Outlook

- Investigation of the effect of different hyper-parameter settings
    - Can they be optimised automatically?
    - Can we find task or dataset invariant heuristics to optimise them?
- More effective use of unlabelled data
    - Can we identify a subset of unlabelled data that better aligns with the current analysis?

Outlook

- Investigation of the effect of different hyper-parameter settings
    - Can they be optimised automatically?
    - Can we find task or dataset invariant heuristics to optimise them?
- More effective use of unlabelled data
    - Can we identify a subset of unlabelled data that better aligns with the current analysis?
- The role of opinionated multi-word hashtag expressions
    - What effect do they have in Sentiment Analysis?

Q & A

Contact:

Thomas Kober
t.kober@sussex.ac.uk

David Weir
d.j.weir@sussex.ac.uk

EM in a nutshell

EM in a nutshell

- Start with an initial model (e.g. trained on the available labelled data)

EM in a nutshell

- Start with an initial model (e.g. trained on the available labelled data)
- Label the unlabelled data with this initial model
    - weighting of unlabelled data important
    - baseline uses a static weight of 0.1
    - We apply a heuristic which weights every unlabelled instance with |labelled data| / |unlabelled data|

EM in a nutshell

- Start with an initial model (e.g. trained on the available labelled data)
- Label the unlabelled data with this initial model
    - weighting of unlabelled data important
    - baseline uses a static weight of 0.1
    - We apply a heuristic which weights every unlabelled instance with |labelled data| / |unlabelled data|
- Retrain model on *all* data

EM in a nutshell

- Start with an initial model (e.g. trained on the available labelled data)
- Label the unlabelled data with this initial model
    - weighting of unlabelled data important
    - baseline uses a static weight of 0.1
    - We apply a heuristic which weights every unlabelled instance with |labelled data| / |unlabelled data|
- Retrain model on *all* data
- Iterate until stopping criterion is met
    - Typically until model parameters converged
    - We stop after 1 iteration (mainly for reasons of running time)

Bernoulli vs. Multinomial

Bernoulli vs. Multinomial

- Bernoulli
    - Explicitely models the absence of a feature
    - Models the number of documents of class c containing feature f

Bernoulli vs. Multinomial

- Bernoulli
  - Explicitely models the absence of a feature
  - Models the number of documents of class c containing feature f
- Multinomial
  - Absence of a feature implicitly modelled in class-conditional probabilities
  - Models the number of times feature f appears in documents of class c

Bernoulli vs. Multinomial

- Bernoulli
  - Explicitely models the absence of a feature
  - Models the number of documents of class c containing feature f
- Multinomial
  - Absence of a feature implicitly modelled in class-conditional probabilities
  - Models the number of times feature f appears in documents of class c
- binary Multinomial
  - Same as Multinomial, but feature counts are capped at 1