INFORMATION EXTRACTION FROM DIGITAL SOCIAL TRACE DATA: APPLICATIONS IN SOCIAL MEDIA AND SCHOLARLY DATA ANALYSIS

BY

SHUBHANSHU MISHRA

DISSERTATION PROPOSAL

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in School of Information Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Assistant Professor Jana Diesner, iSchool, Chair & Director of Research
Associate Professor Vetle I. Torvik, iSchool
Professor Karrie Karahalios, Computer Science
Professor Robert J. Brunner, iSchool & Astronomy

Abstract

This work proposes information extraction (IE) systems for digital social trace data (DSTD). DSTD are traces of digital activity generated as part of social interactions. In this work, a DSTD is a representation of social interactions, which includes temporal dependence as well as node and edge level meta-data. This thesis proposes to examine the utility of DSTD representation for more accurate and interpretable information extraction systems for social networks. The utility of the proposed approaches is demonstrated using examples from the domains of social media and scholarly publishing. For social media data the tasks of opinion extraction and named entity recognition (NER) are studied. For opinion extraction, the usage of actionable labels (such as support / non-support and enthusiastic / passive) are suggested to supplement traditional labels (positive / negative / neutral). This work demonstrates the effectiveness of using active human-in-the-loop learning for opinion extraction, and semi-supervised learning for NER. Furthermore, a single end-to-end multi-task learning model, based on deep learning techniques, is proposed for efficient utilization of labeled corpora for different tasks to improve IE on DSTD. Correlation between meta-data features of tweets and their opinion labels in benchmark corpora are studied to understand biases in opinion labels. Additionally, a technique called Social Communication Temporal Graphs (SCTG) is proposed for visualizing DSTD. The extraction and visualization of conceptual novelty and expertise in scholarly publications is also investigated. This thesis proposes the development of open source tools for information extraction applied to DSTD.

To my wonderful family.

Table of Contents

List of Tables . . . . . **vi**

List of Figures . . . . . **vii**

List of Abbreviations . . . . . **viii**

Chapter 1 Introduction **1**
1.1 Digital Social Trace Data (DSTD) 1
1.1.1 Social media data as DST 2
1.1.2 Scholarly publishing data as DST 2
1.1.3 Other examples of DSTD 3
1.1.4 Information extraction on DSTD 3
1.2 Information Extraction Tasks 4
1.2.1 Sentiment prediction 4
1.2.2 Named entity recognition, classification and linking 5
1.2.3 Concept extraction and mapping 5
1.2.4 Other tasks 5
1.3 Existing challenges in IE research for DSTD 6
1.4 Existing approaches 7
1.5 Research Questions 7
1.6 Proposed methods and solutions 8

Chapter 2 What information to extract? **10**
2.1 Socially relevant sentiment labels 10
2.1.1 Background 10
2.1.2 Dataset 10
2.1.3 Analysis 11
2.2 Meta data association with sentiment 12
2.2.1 Dataset 13
2.2.2 Analysis 14

Chapter 3 How to extract information? **16**
3.1 Incremental learning of sentiment with human in the loop 18
3.1.1 Model 18
3.1.2 Analysis 19
3.2 Semi-supervised entity recognition 20
3.2.1 Background 20
3.2.2 Dataset 20
3.2.3 Analysis 20
3.3 Deep multi task multi dataset learning 22
3.3.1 Deep Learning for Information Extraction 22
3.3.2 Deep multi-task learning 22

3.3.3 Background . . . 23
3.3.4 Dataset . . . 23
3.3.5 Model . . . 23

Chapter 4 Applications and presentation of extracted information . . . 26
4.1 SCTG: Social Communications Temporal Graph A novel approach to visualize temporal communication graphs from social data . . . 26
4.1.1 Background . . . 26
4.1.2 Components . . . 27
4.1.3 Applications . . . 28
4.2 Quantifying Conceptual Novelty in the Biomedical Literature . . . 28
4.2.1 Data . . . 29
4.2.2 Analysis . . . 29

Chapter 5 Conclusion . . . 33
5.1 Limitations . . . 33

Chapter 6 Thesis timeline . . . 35

References . . . 36

List of Tables

2.1 Inter-annotator agreements for various sentinets datasets 11
2.2 Label distribution across various sentiment datasets 14

3.1 Prediction accuracy depending on training algorithm and feature sets 19
3.2 Change in F1 score for the NER classifier on the development dataset on incremental addition of different types of features (from left to right). ST refers to submitted solution, BL refers to baseline solution provided by the organizers. Bold values are the best scores across classifiers. 21
3.3 List of datasets used for social media analysis 24
3.4 Description of datasets for named entity recognition on Tweets 24

List of Figures

1.1 Illustration of digital social trace data . . . 3

2.1 Distribution of labels in the Sentinets dataset . . . 11
2.2 Learning curves of logistic regression model on Sentinets dataset . . . 12
2.3 Mean corrrelation of tweet meta-data with annotated sentiments on user timelines . . . 14
2.4 Mean corrrelation of tweet meta-data with annotated sentiments on all data . . . 15

3.1 Neural network uncertainty . . . 17
3.2 Model for training sentiment using human-in-the-loop incremental learning . . . 19
3.3 Human in the loop application interface . . . 19
3.4 Frequency of named entity types in training, development, and test datasets . . . 20
3.5 Model architecture . . . 21
3.6 Encoder decoder framework . . . 23
3.7 Continual learning using EWC . . . 25

4.1 Visualization of conversation growth in a Facebook course group . . . 27
4.2 Growth of MEDLINE . . . 29
4.3 Temporal profile of HIV . . . 31
4.4 Correlation of novelty with citations . . . 32

6.1 Thesis timeline . . . 35

List of Abbreviations

DSTD	Digital Social Trace Data
IE	Information Extraction
ML	Machine Learning
DNN	Deep Neural Network
SSL	Semi-Supervised Learning
MTL	Multi Task Learning
HITL	Human In The Loop
POS	Part-Of-Speech
NER	Named Entity Recognition
NEL	Named Entity Linking
NERD	Named Entity Recognition and Disambiguation
NERC	Named Entity Recognition and Classification

Chapter 1

Introduction

Information Extraction (IE) deals with the generation of structured output from unstructured data [Sarawagi, 2008]. In the domain of text data, IE is used to extract salient topics or concepts, with a particular set of application being named entity recognition, classification, disambiguation, and linking [Nadeau and Sekine, 2007, Tjong Kim Sang and De Meulder, 2003]. Other popular IE tasks are relation extraction and classification [Mintz et al., 2009], and automatic knowledge base (KB) construction[Socher et al., 2013]. Most of these tasks usually rely on features from natural language processing tasks like part of speech (POS) tagging, phrase parsing, dependency parsing, and co-reference resolution [Sarawagi, 2008]. Finally, sentiment (or opinion) extraction [Pang and Lee, 2008], is also considered an instance of an IE task. Here sentiment is merely a measure of the perceived or intended sentiment in the text [Pang and Lee, 2008].

Social scientists are interested in the evolution of social systems. The current rise in usage of social media platforms like Twitter, Facebook, Reddit, gives them a suitable benchmark to test existing social science theories[Kosinski et al., 2015, Wilson et al., 2012, Miller, 2011, Lazer et al., 2009, Kwak et al., 2010]. An important aspect of social science research on social media data is the analysis of social media text and meta-data. However, existing IE and NLP systems have been found to perform poorly on social media text (also known as Noisy User-generated Text (NUT) [Baldwin et al., 2015]) compared to text from news corpora (for which most IE systems were made). A major reason for this poor performance is the usage of non-traditional vocabulary, word forms, and short message length [Eisenstein, 2013].

1.1 Digital Social Trace Data (DSTD)

In order to make IE accessible to social scientists and to make it applicable to a broad range of data sets, we define an abstract concept of Digital Social Trace Data (based on the concept of digital social trace data [Diesner and Chin, 2015] and digital trace data [Howison et al., 2000]) or DSTD. DSTD are digital activity traces generated by individuals as part of a social interactions, such as interactions occur on social media websites like Twitter, Facebook; or in scientific publications. A DSTD has the following properties:

- Temporal distribution of the data
- Presence of connection between various data items
- Associated meta-data for each multiple parts of the data.

DSTD are very similar to heterogeneous information networks (HINs) [Sun and Han, 2012], and temporal networks [Holme and Saramäki, 2012]. While HINs represent information network as a "typed semi-structured heterogeneous networks" [Sun and Han, 2012], where nodes and edges are typed. On the other hand, DSTD are focused on simplifying this relationship into three core components mentioned above. Similarly, temporal networks only capture the temporal evolution of the network and are commonly discussed in the homogeneous network setting. Again, DSTD expand on this definition by including meta-data attributes. Finally, DSTD are described in a social science setting as opposed to the graph theoretic setting. This will ensure effective communication of DSTD to the social science community.

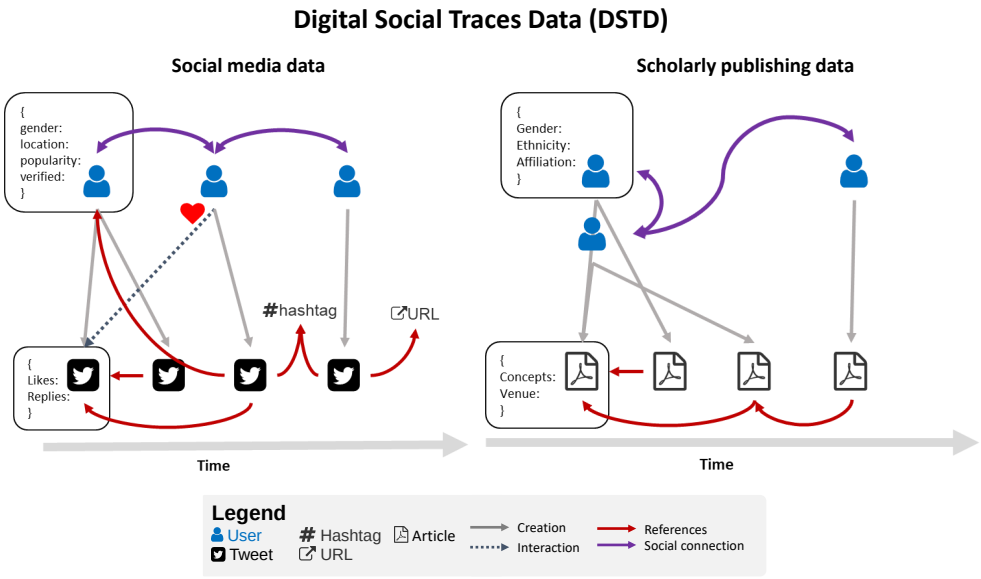
1.1.1 Social media data as DST

An example of a DST will be an individual's Twitter feed, or a collection of all the Tweets about *Hurricane Irma*. Both these dataset will have a temporal distribution as the Tweets are generated in a temporal order. Furthermore, the interaction of tweets with each other satisfies the connection presence. This can be seen when other users re-tweet a given tweet, or reply to it. Finally, tweets carry a large amount of meta-data both about its content as well as about its creator, e.g. tweets may be tagged by a geo-location, or contain a URL. The meta-data about the user of the tweet usually contains the user's location, number of followers, friends, and current number of posted-tweets. Figure 1.1 gives an illustration of DSTD for social media and scholarly publishing data. Furthermore, social dialogs can be identified as DSTD.

1.1.2 Scholarly publishing data as DST

Similar to social media data, scholarly publishing data can also be considered an instance of DSTD. In the most common scenario, scholarly publication datasets consists of articles, their associated meta-data, and the citation network. However, this dataset can be expanded to include authors and publication data, to convert the dataset into an instance of DST. Figure 1.1 shows how scholarly publishing data is an instance of DST.

Figure 1.1: Illustration of digital social trace data



1.1.3 Other examples of DSTD

Many datasets can be modeled as DSTD by ensuring certain properties of the dataset are available–such as time-stamp of individual activities; connection between individuals and items; meta-data associated with each individual and item.

1.1.4 Information extraction on DSTD

What? Every IE requires identification of what information to extract. Earlier research in IE was focused on improving search engine results quality or question answering systems. However, for social scientists the relevant questions are different. Hence, the labels and kind of extracted information also needs to change. For example the popular task of sentiment analysis uses labels positive/negative/neutral for sentiment, whereas the meaning of these terms is not clear from the perspective of social science theory. Herein, a more actionable label such as supportive or non-supportive, can make it easier for the researcher to map the output of the IE system to their existing knowledge. Hence, a major question to ask when doing IE for social scientists, is what information should be extracted?

How? Additionally, a major challenge in building any IE system is to design evaluation data to assess the quality of the extracted information. Furthermore, with the advancements in supervised learning algorithms, there is a demand for building larger and higher quality training corpora for any IE system. In the domain of social media DSTD, this training data is very scarce, annotated with varying guidelines, and is biased

to certain time spans, domains, or geo-locations. This challenge promotes an investigation into how to efficiently extract this information? Specifically, how best can the existing annotated sources be utilized to bring the social media IE systems on par with existing IE systems.

Applications Next, once information is extracted it needs to be made accessible to the social science community. Specifically, the presentation of the extracted information should follow the structure of the information, i.e. temporal, connected, and meta-data enriched. However, existing IE systems have not focused on these aspects, causing difficulty in interpreting the extracted information at an aggregate level. Here lies the third challenge, how to make the extracted information more accessible.

Finally, scholarly publishing data is also an example of DSTD. In scholarly publishing publications are temporally ordered, have connections between each other via-citations and co-authorship networks, and are tagged with additional meta-data such as concepts, publication venue, author's prior publication count, author's h-index. Using these data additional information can be extracted such as an overall conceptual novelty of an article [Mishra and Torvik, 2016].

1.2 Information Extraction Tasks

Here we discuss two broad category of tasks:

- **Text classification**: This includes opinion mining, geo-location prediction
- **Text tagging**: This includes named entity extraction, part of speech tagging, chunking, super-sense tagging, dependency parsing.

Below we will discuss each task in detail and describe their utility towards extracting information from social media.

1.2.1 Sentiment prediction

Sentiment prediction is typically modeled as a text classification problem such the that sentiment $y \in labels$ is dependent on the text based features of the data and is modeled as $p(y|X)$, where X are features derived from the text. The most commonly used labels for sentiment are *positive* and *negative*, with the optional inclusion of *neutral*. However, a consistent meaning of these labels is not adopted across datasets. Furthermore, sentiment itself is a highly subjective quantity which can be described using the state of the content author, as well as the state of the receiver of the sentiment. This can be demonstrated using an example. Consider the tweet *"Roger Federer killed this, Nadals sucks."* [Mishra et al., 2014]. In the presented tweet, suppose

the author is a fan of Federer, and Federer won the match with Nadal. Here, the author is most likely to show their support for Federer and their opposition of Nadal. However, if it was Nadal who won the match, and the author is a Federer fan, then the author shows their dislike towards Federer (or is not in support of Federer's game in that match), as well as their dislike towards Nadal (overall). Similarly, if the author was a Nadal fan, and the game was won by Federer, they would have surely shown their dislike towards Nadal's game while appreciating Federer. The example serves as a demonstration of the subtle nuances of sentiment analysis which are not incorporated in most application scenarios. These approaches are often studied under aspect based sentiment analysis[Pontiki et al., 2015].

1.2.2 Named entity recognition, classification and linking

Named entity recognition is the identification of named entities in text. Here named entities are single or multi word units of text, e.g. *Barack Hussein Obama*. Furthermore, named entities are usually tagged with the type of entity, e.g. *Barack Hussein Obama* is an entity of type *person*, it can also be an entity of type *political figure* given the context of the text. This is commonly referred to as named entity classification. A common task in IE is to perform named entity recognition and classification together (NERC). The output of NERC can be further enhanced by linking named entities to existing knowledge bases such as Wikipedia, or Wikidata. This task is commonly referred to as named entity linking or disambiguation (NELD). Named entities can be utilized for improving search query results, building better question answering systems, as well as for identifying the target of sentiment in text.

1.2.3 Concept extraction and mapping

Concept extraction is a task aimed at identifying key concepts in text data. In the domain of scholarly publishing, concepts are defined for a domain, e.g. Computer Science, and the goal is to extract concepts from an article and map them to a domain specific lexicon e.g. Medical Subject Headings (MeSH).

1.2.4 Other tasks

Additional tasks are commonly studied under information extraction. Some of these such as part of speech tagging, chunking, super-sense tagging can be utilized as pre-processing steps for the earlier mentioned tasks, while other tasks such as geo-location prediction [Han et al., 2016], rumor classification[Zubiaga et al., 2016a, Zubiaga et al., 2016b], author-profiling [Rosso et al., 2016] are studied in isolation.

This thesis proposal is divided into six chapters. Chapter 1 gives an overview of the information extraction, and the definition of digital social trace data, it also identifies the main goals of this thesis. Chapter

2 focuses on identifying what information to extract for social science domains, with focus on opinion extraction at user level and named entity recognition. Chapter 3 described the methodological frameworks which are appropriate for extracting information from DSTD, such as active human-in-the-loop learning, semi-supervised learning, and multi-task learning. Chapter 4 provides examples on novel applications of the extracted information such as a new visualization framework for DSTD. Additionally, examples are provided from IE on scholarly publication data such as publication and author level conceptual novelty and expertise in biomedical literature. Chapter 5 deals with the limitations and concluding remarks. Chapter 6 describes the timeline of the thesis work.

1.3 Existing challenges in IE research for DSTD

Information extraction in DSTD, especially social media, suffers from major challenges because of a more diverse community of users who post content in different languages, dialects, and domains. Since many IE systems are based on supervised machine learning techniques, they suffer from the issue of domain adaptation. Consider opinion mining or sentiment analysis as an IE task, earlier research has shown that sentiment in social media is more nuanced, and harder to predict compared to newswire or review corpora [Mishra et al., 2014, Mishra et al., 2015, Maynard et al., 2012, Aue and Gamon, 2005].

A more practical issue in performing IE on social media DTS data is the lack of standardized annotated corpora similar in quality and scale of Penn Tree-Bank, Universal Dependencies[Nivre et al., 2016], Movie Review corpus[Pang and Lee, 2008]. In recent years efforts have been made to construct such corpora e.g. sentiment analysis [Nakov et al., 2016a, Nakov et al., 2016b], named entity recognition [Ritter et al., 2011, Derczynski et al., 2016, Baldwin et al., 2015], and part of speech tagging [Derczynski et al., 2013, Owoputi et al., 2013]. However, many of these datasets suffer from varying tokenization issues, tag annotation discrepancies, and disproportionate tag distribution [Mishra and Torvik, 2016]. Furthermore, many of the IE tasks are constructed as pipeline of tasks e.g. NER systems usually pre-process the text with tokenization, part of speech tagging, and noun phrase chunking, before training a NER model. Similarly, many aspect based sentiment analysis systems require the extraction of named entities from text and then assign a sentiment to each entity. However, in case of social media data the error in the pre-processing models are very likely to propagate in the training of the final model.

In terms of presenting the information extraction results from DSTD, we still stick to one of its views, choosing either to show its temporal, networked, or meta-data based aspect. Many approaches combine the meta-data aspect with the networked or temporal aspects, but this still leaves an important gap in seeing

the DSTD for what it is.

1.4 Existing approaches

In recent years scholars have proposed using advanced machine learning (ML) methods for improving IE systems. A common objective of these advanced methods is reducing the amount of human effort in annotating the training data. In the field of sentiment analysis author have proposed using distant supervision [Mintz et al., 2009] for training models using emoticons present in tweets [Go et al., 2009, Felbo et al., 2017]. Distant supervision utilizes noisy labels to train the model on large unlabeled data. Distant supervision has also been applied for doing POS tagging of tweets [Plank et al., 2014].

Similarly, active human-in-the-loop learning also known as interactive machine learning (iML) allows the model to collect high quality training data by collecting annotations based on the model's uncertainty [Settles, 2009]. This is important as usually the cost of annotating data especially those with structured output spaces is high [Settles et al., 2008]. Active human-in-the-loop learning has been successfully applied for sentiment analysis [Mishra and Torvik, 2016] and sequence labeling tasks such as [Settles and Craven, 2008].

Another popular approach for efficiently training machine learning systems is using Semi-Supervised Learning (SSL) [Chapelle et al., 2006, Zhu, 2008]. This approach uses unlabeled data together with labeled data to learn the model under certain assumptions of the distribution of the data. It has been successfully applied to named entity recognition of tweets [Mishra and Torvik, 2016]

1.5 Research Questions

The main research questions of this thesis are as follows:

RQ1 *What information to extract?* This addresses the need to identify what information is useful for social science research.

RQ2 *How to efficiently extract information?* This addresses the importance of using machine learning algorithm which are more suitable to the nature of DSTD.

RQ3 *How can the extracted information be presented and utilized?* This addresses the need for new visualization and presentation interfaces to make the extracted information from DSTD more accessible to social science research community.

1.6 Proposed methods and solutions

In order to solve each research question we propose the following solutions:

RQ1 Suggest an alternative orthogonal set of labels and annotated data which identifies if a tweet supports or opposes the cause and if it conveys an author's enthusiasm or passiveness towards the cause. Extract bias towards user and tweet meta-data in sentiment annotated corpora.

RQ2 Use active human-in-the-loop learning, semi-supervised learning, and multi-task learning for improving sentiment extraction and named entity recognition in tweets.

RQ3 Present a novel visualization framework for DSTD which allows presenting temporal, network, and meta-data aspects of the corpus. Show applications of extracting novelty and expertise from large biomedical corpora.

The above mentioned approaches can be summarized into the following goals:

- Move away from default labels positive versus negative, to task specific labels e.g. Enthusiastic versus Passive, and Supportive versus Non-supportive. This will help social scientists evaluate the models trained on these datasets by grounding them in prior literature.
- Allow the model to learn over time online learning of classifier using online modeling techniques as well as data augmentation techniques like updatable gazetteers. Effective for NER in tweets.
- Visualize the social network aspect of the text data Visualization.
- Meta data (such as users network and post interactions) can be used for improving the classification accuracy of existing models. How are meta features correlated with sentiment labels?
- Show the benefit of using multi-task learning approaches for tasks where training data is sparse by utilizing training data for similar tasks, e.g. sentiment prediction, POS, NER.

Finally, the thesis proposes to have the following output:

- An annotated set of data for alternative opinion labels.
- GUI based tool to allow online learning of text classification and sequence labeling models with data augmentation.
- Visualizing the network structure of social conversation using a temporal network visualization which can be modified to show user as well as post level attributes.

- List of meta features which can improve text classification tasks.
- Principled approach and ready to use tool for multi task learning of supervised models which use information from differently annotated corpora.
  - Does order of learning tasks matter?
  - Different between joint training versus training incrementally?
  - Computational cost of using a multi task model
- Consolidate the existing corpora for learning from social media data and annotate for multiple tasks and map to universal dependencies data. [low priority]

Chapter 2

What information to extract?

2.1 Socially relevant sentiment labels

This work will build up on our prior work on sentiment analysis of social causes on Twitter [Mishra et al., 2014]. Socially relevant sentiment labels in our case refer to annotation labels which reference an intent or action of the author as conveyed by their text. For this purpose we collected data on three social issues namely, *Cyberbullying*, *LGBT*, and *Concussions (CTE) in sports*. These topics were chosen because of their controversial nature and a broad scope of demographic interest. Our research question for this research was oriented towards identification of users, who are likely to support or be against (non-support) one of these social causes. Additionally, we wanted to identify the nature of support or non-support. We did this by identifying support or non-support across the spectrum of the author being enthusiastic or passive in the text. Existing work on sentiment analysis usually deals with *positive*, *negative*, and sometimes *neutral* labeling of the text. In many cases these labels don't convey enough information to the consumer of these labels about the intent of the author of the text. Thereby, our task setting resulted in defining the user sentiment across two orthogonal dimensions namely **Support** and **Enthusiasm**. This dataset will be referred to as **Sentinets** dataset from now onwards.

2.1.1 Background

Opinion mining is [Pang and Lee, 2008] the task of extracting user opinions from data. It may help in identifying datasets which express a specific type of opinion as well as the degree of that opinion. Opinion mining has been a very prominent topic of interest in social media research because of its application to predicting elections[Tumasjan et al., 2010] and stock prices [Bollen et al., 2011].

2.1.2 Dataset

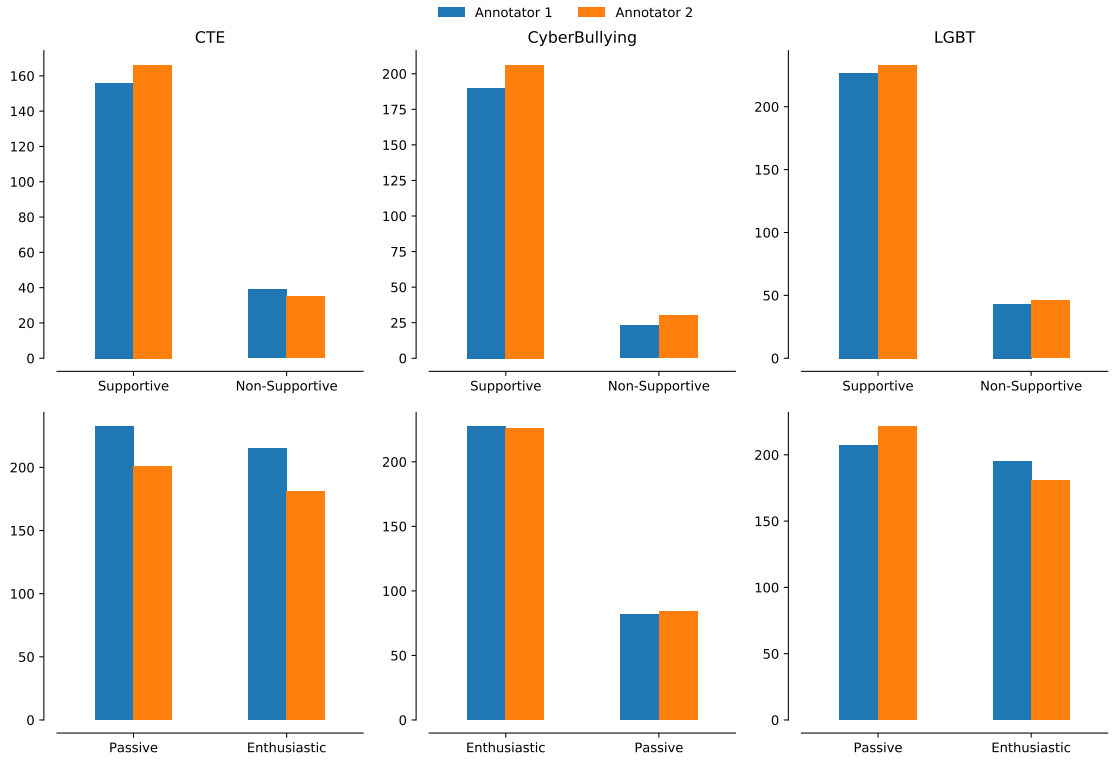
For each of the topics we downloaded recent tweets from twitter in 2013 and created a detailed codebook for annotation of each of these tweets across these sentiment dimensions. Two annotators who were well

Table 2.1: Inter-annotator agreements for various sentinets datasets

		κ	N
CTE	**support**	0.92	165
	enthusiasm	0.91	379
CyberBullying	**support**	1.00	209
	enthusiasm	0.86	309
LGBT	**support**	0.89	257
	enthusiasm	0.87	395

versed with the codebook annotated the tweets and reached a high inter-annotator agreement as shown in 2.1. The distribution of labels by each annotator is shown in 2.1.

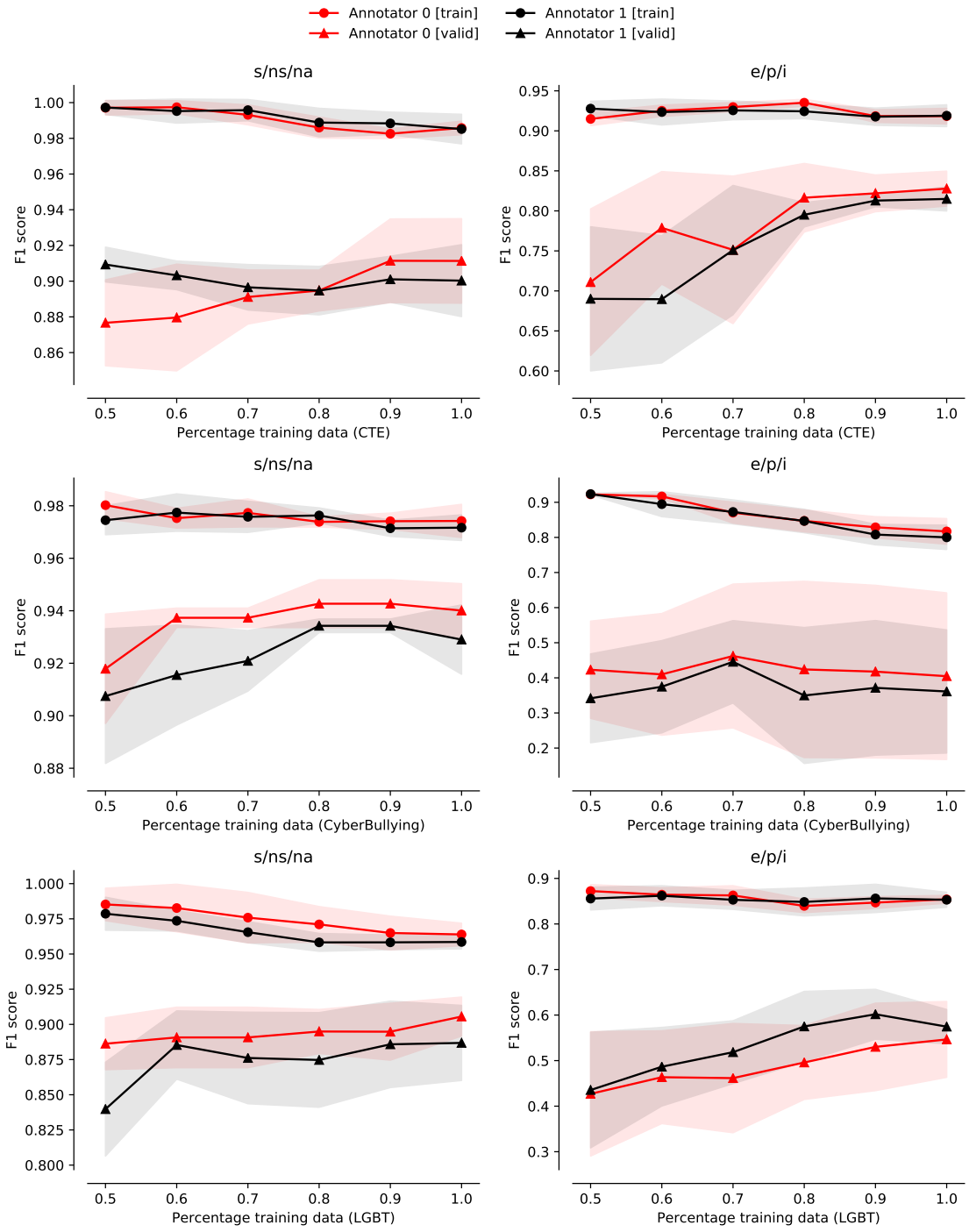
Figure 2.1: Distribution of labels in the Sentinets dataset



2.1.3 Analysis

Finally, we train a five fold cross validated logistic regression model which with just unigram token features which performs very well on the datasets resulting in $\approx$ 90% F1 score on the test splits for the support dimension. The model learning curves are shown in 2.2. The performance on enthusiasm dimensions can be improved using additional features such as POS tags, and sentiment lexicon features.

Figure 2.2: Learning curves of logistic regression model on Sentinets dataset



2.2 Meta data association with sentiment

Social media data is usually more than text. Specifically, for sentiment analysis tasks the social media data contains much richer attributes such as the user's popularity (measured in number of friends, followers),

user's activity level (number of posts, communities the user is part of), and the post specific attributes (number of shares, favorites, and replies). These attributes can be useful in understand the social context of these messages. In this part of our analysis we use six frequently used pre-annotated twitter sentiment corpora and study the association of various sentiment labels with each of the attributes mention. Finally, we describe a meta-data enhanced classifier which can use the user and message attributes as priors to enhance the sentiment predictions.

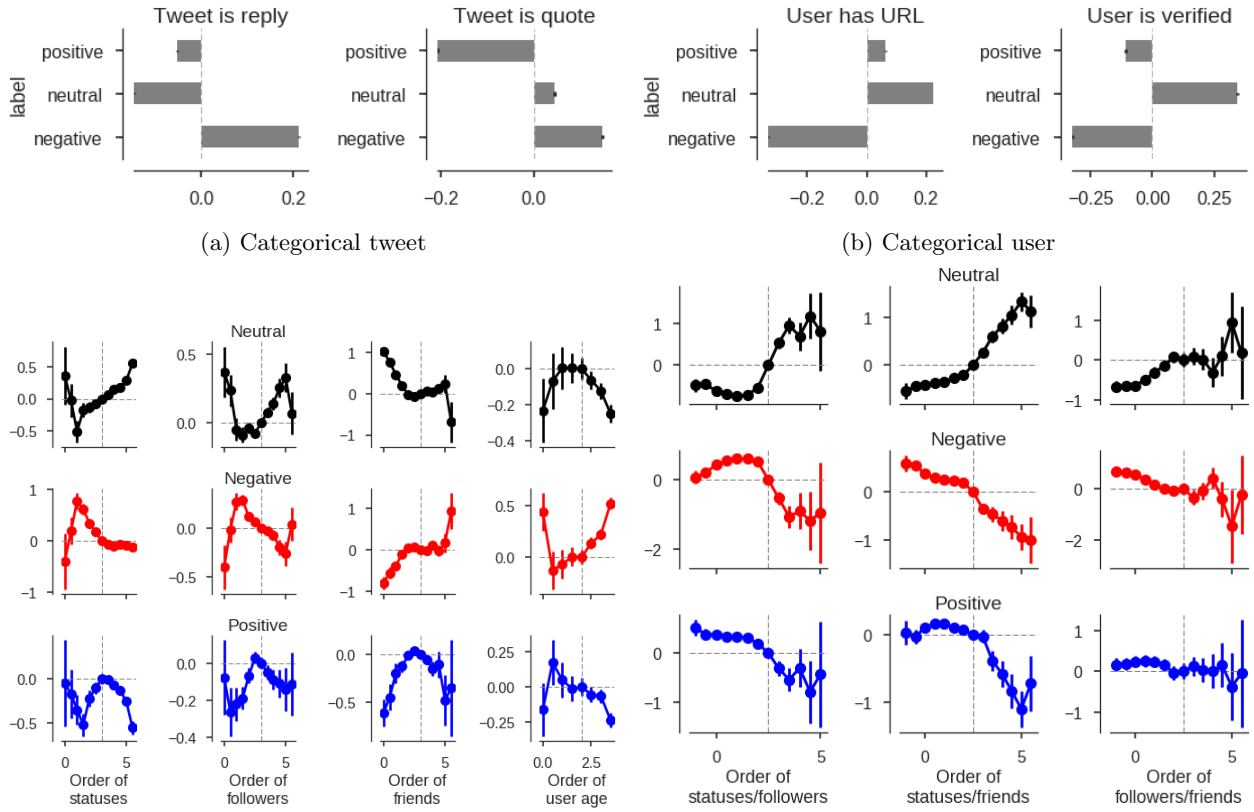
2.2.1 Dataset

Most existing sentiment datasets label the data into three classes, namelynegative, neutral, and positive. We use the same set of labels for our analysis, and only consider datasets which have their tweets annotated with those labels. Additionally, we also consider a different set of binary class labels to identify if tweets are opinionated (either positive or negative) or non-opinionated (neutral). Furthermore, we require that for a dataset to be selected as part of our analysis has tweet IDs for each tweet label. This is important for collecting the user and tweet level metadata using the Twitter API . Finally, to infer any significant relationship between metadata and sentiment labels, we want to avoid any data specific biases in annotation and tweet distribution. We address this bias mitigation by using sentiment labeled datasets from a varied time-periods, on different topics, and that have different annotation schemas i.e. different annotation guidelines and annotation interfaces were used for each data annotation. Using this methodology, we should be able to infer the average relationship between tweet metadata and sentiment labels after pooling the selected eligible datasets. Based on our above-mentioned criterion, we identify six high quality, publicly available datasets for our analysis. The first dataset (referred as SemEval) is from the recurring Twitter sentiment classification task of SemEval [Nakov et al., 2016b, Nakov et al., 2016b], and includes all training, development and test data from 2013 throughout 2016. We only consider the data for the tasks where the goal was to classify tweet sentiment as either negative, neutral, or positive. The second dataset is a publicly available dataset of large collection of multilingual tweets from European countries, from the study of Mozetic and colleagues [Mozetič et al., 2016]. We only work with the English tweets from this dataset. This dataset is available on the CLARIN data repository and therefor referred to as Clarin. The next two datasets namely, Airline and GOP generated by the Crowdflower platform and hosted on Kaggle , include crowd sourced sentiment annotations for tweets about various Airlines and the first GOP debate of 2016. The final two datasets come from Saif and colleagues [Saif et al., 2013] and are about the Obama-McCain debate (referred as Obama) and healthcare (referred as Healthcare).

The dataset distribution is described in Table 2.2

Table 2.2: Label distribution across various sentiment datasets

Dataset	Train			Development			Test			Total
Labels	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	
Airline	5,515	1,843	1,467	12,258	205	163	1,532	512	408	12,258
Clarin	11,485	19,418	13,496	61,667	2,158	1,500	3,191	5,394	3,749	61,667
GOP	4,230	1,818	1,173	10,030	202	130	1,175	505	326	10,030
Healthcare	834	378	321	2,131	42	36	232	106	89	2,131
Obama	715	707	455	2,608	79	50	199	197	126	2,608
SemEval	4,313	13,031	11,405	39,931	1,448	1,268	1,198	3,620	3,169	39,931



(a) Categorical tweet

(b) Categorical user

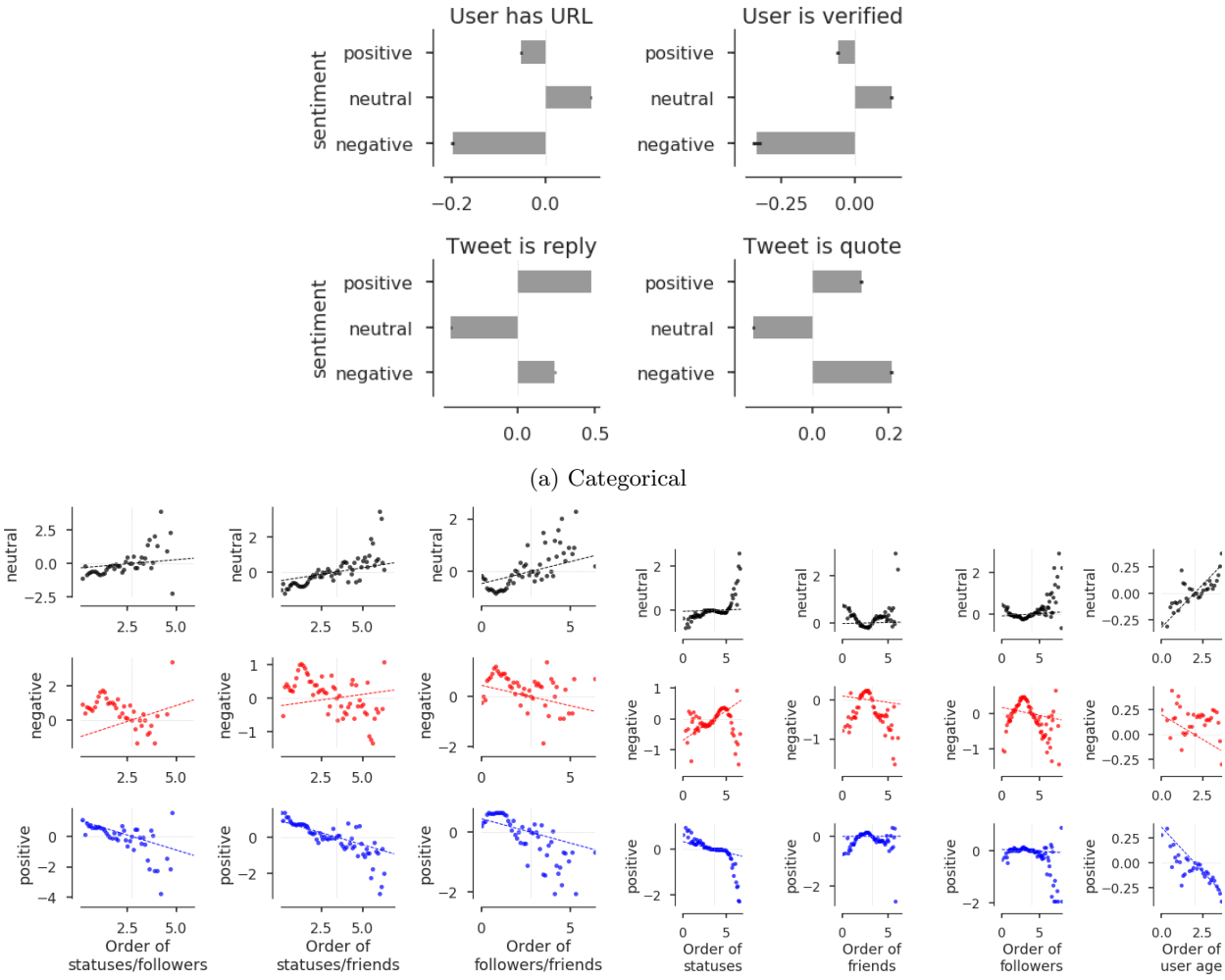
(c) Numeric

(d) Numeric ratios

Figure 2.3: Mean corrrelation of tweet meta-data with annotated sentiments on user timelines

2.2.2 Analysis

We analyze the correlation between different features of the user and tweets to the sentiment labels. In order to ensure that these correlations are consistent for all the tweets of the users in our dataset, we additionally collect 200 most recent tweets for each user and label them with sentiment using an off the shelf classifier. Our analysis of the meta-data correlations is presented in figure 2.3 and figure 2.4. Overall, we find that a majority of trends are consistent across the data.



(a) Categorical

(b) Numeric ratio

(c) Numeric

Figure 2.4: Mean corrrelation of tweet meta-data with annotated sentiments on all data

Chapter 3

How to extract information?

This chapter deals with the methodological approach for performing IE on DSTD. DSTD are challenging from machine learning point of view. One major property of DSTD is the temporal change of data. This means that the data samples are not i.i.d. (independent and identically distributed). In most cases the data samples are dependent on the data produced in the past, e.g. the meaning of words phrases changes over time. Furthermore, data items are also correlated with each other because of the network connections between the data items, e.g. tweets produced by the same author will usually follow the same linguistic structure. Finally, the very nature of DSTD causes the user to never observe i.i.d samples from the true distributions. All three challenges bring forth a limitation of using traditional machine learning techniques as most of them make the i.i.d. assumption about the data. However, in most cases the i.i.d. assumptions are considered valid because of reduced model complexity and computational costs. Herein, lies the opportunity to improve these IE systems by using some recent advancements in building modular machine learning models, namely deep learning based approaches, which are can be easily trained in end to end manner via back-propagation and allow greater flexibility in modeling the data generation process. In order to build better IE systems for DSTD, we need efficient ways of collecting labeled data, utilizing unlabeled data, and labeled data from different tasks and domains. We utilize the concept of model uncertainty (as demonstrated in Figure 3.1) and continual learning of the model (as demonstrated in Figure 3.7) to implement some of our solutions.

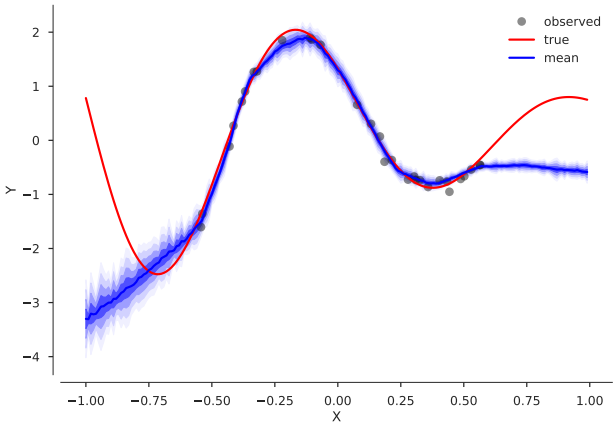


Figure 3.1: Capturing uncertainty in machine learning models using dropout. Red curve is the true function, from which grey samples are observed. Using a 2 hidden layer multi-layer perceptron with 20 units with ReLU (Rectified Linear Unit) non-linearity in first hidden layer, 20 units with sigmoid non-linearity in next hidden layer, and 0.01 dropout probability. The blue line shows the mean prediction by the model while the blue bands show half standard deviation on each side.

3.1 Incremental learning of sentiment with human in the loop

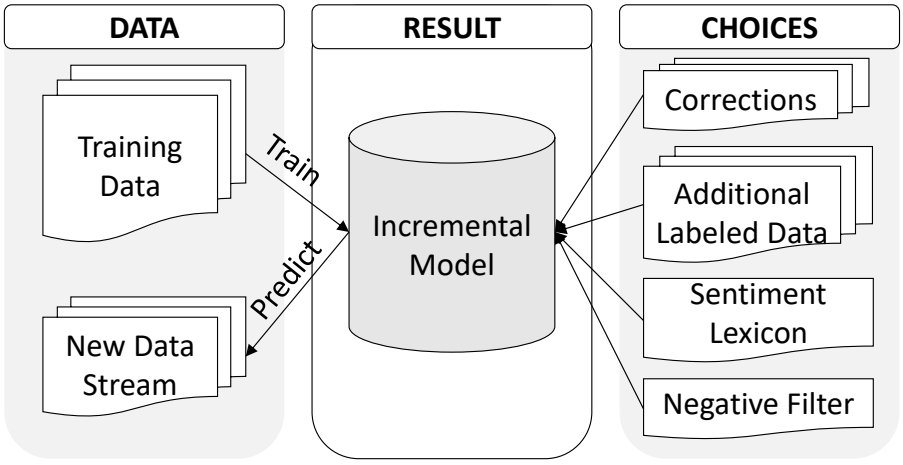
Many sentiment analysis tools apply previously trained models with fixed features and weights to new and unseen data; hoping to obtain accuracy rates similar to those obtained when evaluating the models via k-fold cross-validation. However, trained models can be skewed towards the genre and domain of the training data. Moreover, as language use and the perception of products might change over time, such static models might need to be updated by a) relabeling some prediction results and/or b) adding new labeled instances for learning, and considering either one modification for model updating. This step can be realized via incremental learning, which keeps computational costs low as it updates a model based on changed labels or added instances [Bottou, 2010, Bottou, 1991]. Another issue with sentiment analysis is that several solutions rely on predefined lexicons for mapping tokens from the text data to sentiment categories, or as an additional feature for learning (some cutting edge solutions use bag-of-word approaches that consider more context [Go et al., 2009, Mohammad et al., 2013], or word vector-based deep learning [Mikolov et al., 2013, Pennington et al., 2014] instead). Due to their intended general applicability, existing resources though convenient to use can lead to errors when general terms have different connotations in specific domains. Prior research has shown that sentiment prediction accuracy can be improved by adjusting these lexical resources to a new dataset and domain . This adjustment entails removing false positives from lexicons and adding in false negatives. We have been addressing both of these issues by building a free and open tool (Sentiment Analysis and Incremental Learning, short SAIL, https://github.com/uiuc-ischool-scanr/SAIL) that allows for a) incremental learning and b) adjusting lexical resources (positive and negative filters) (overview shown in Figure 3.2). SAILs baseline model is trained on SemEval data [Nakov et al., 2016a]. Users can also train a model from scratch using their own annotated data and even their own categories.

3.1.1 Model

Data Pre-processing

For each tweet, the content of hashtags, URLs, mentions, emoticons and double quotes was converted into binary mentions (_HASH, _URL, _MENTION, _EMO, _DQ). Each tweet was converted into a vector with the following features: a) Meta: Count of hashtags, emoticons, URLs, mentions, double quotes; b) POS: Count of parts of speech using the ark-tweet-nlp tool [Owoputi et al., 2013, Owoputi et al., 2012]; c) Word: Presence of the top 10,000 unigram and bigram with at least three occurrences; d) Sentiment lexicon: Count of positive and negative words matching a widely used sentiment lexicon [Wilson et al., 2005], which the user can edit; e) Negative filter: A user generated list of words, hashtags and usernames that may represent

Figure 3.2: Model for training sentiment using human-in-the-loop incremental learning



(a) Model loading

Figure 3.3: Human in the loop application interface

false positives with respect to the sentiment lexicon, and hence are omitted from consideration for feature d).

Our model interface is defined in 3.3

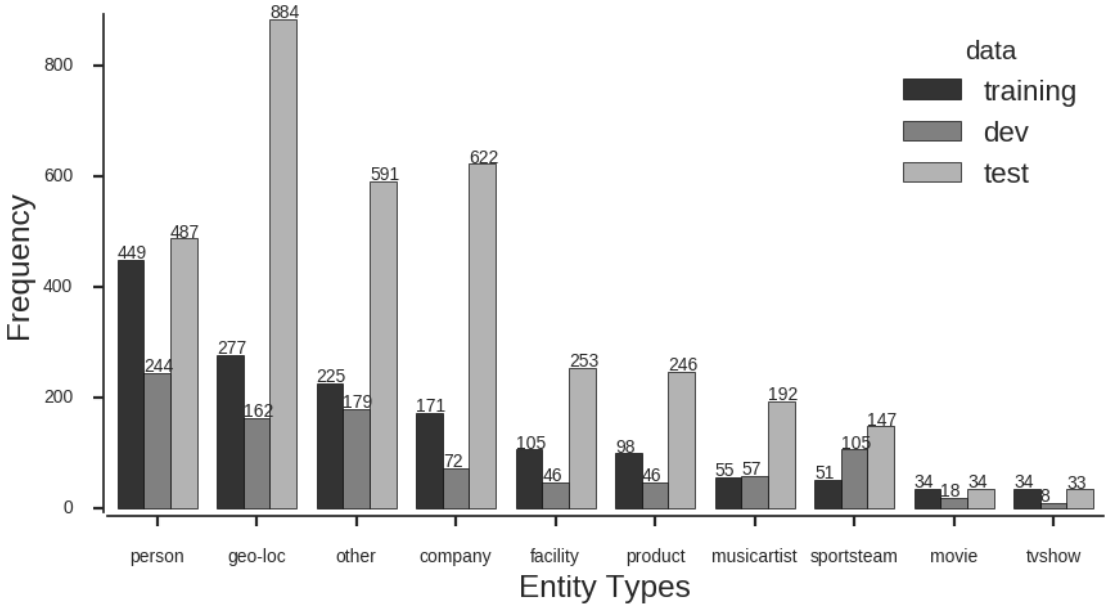
3.1.2 Analysis

We use the SemEval dataset [Nakov et al., 2016b] partitioned into years. The comparison shows that SVM (as implemented in Weka [9]) is only outperformed by SGD (by about 0.9%) when using a large amount of tokens for the word feature Table 3.1.

Table 3.1: Prediction accuracy depending on training algorithm and feature sets

Features considered			**Accuracy (F1)**	
Meta	**POS**	**Word**	**SVM**	**SGD**
X	X		70.50%	70.40%
X	X	X (N=2K)	85.70%	85.60%
X	X	X (N=20K)	86.60%	87.50%

Figure 3.4: Frequency of named entity types in training, development, and test datasets



3.2 Semi-supervised entity recognition

3.2.1 Background

Semi-supervised learning can be useful for many tasks where we have large amounts of unlabeled data as well as some labeled data. The motivation behind using semi-supervised learning is that both these datasets can be utilized efficiently to build a more generalizable classifier than simply using the labeled data. The key idea is to use the unlabeled data as some kind of guiding prior for model.

3.2.2 Dataset

The dataset labels are described in Fig. 3.4. This dataset is referred to as WNUT16-NER and is described in the paper [Han et al., 2016].

3.2.3 Analysis

Our model architecture is shown in 3.5. The results of training the model incrementally on the dataset with various features are shown in Table 3.2

Figure 3.5: Model architecture

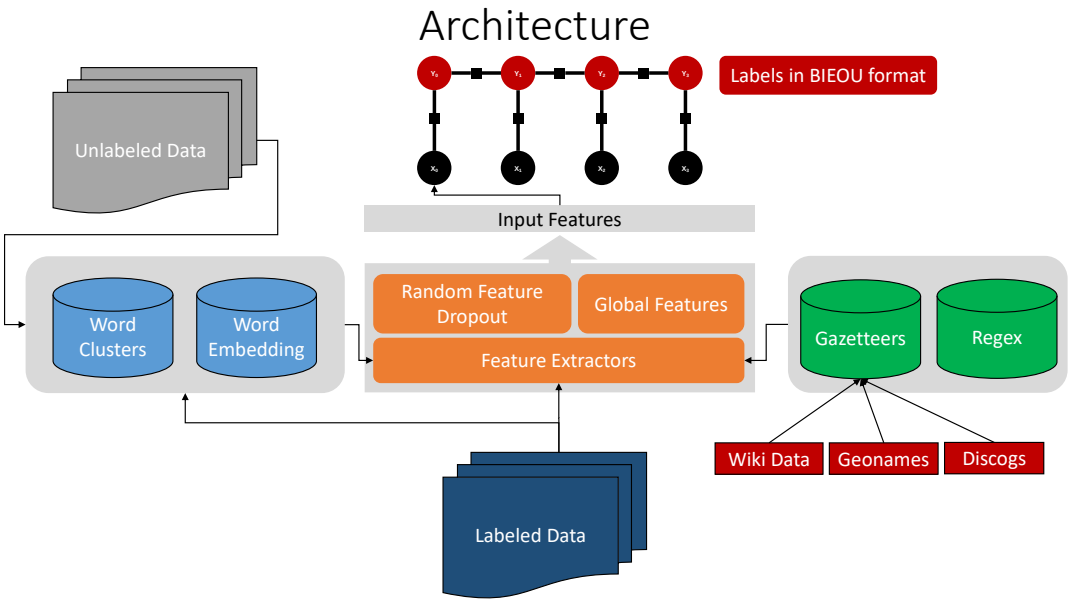


Table 3.2: Change in F1 score for the NER classifier on the development dataset on incremental addition of different types of features (from left to right). ST refers to submitted solution, BL refers to baseline solution provided by the organizers. Bold values are the best scores across classifiers.

Features	RF	+GZ	+WR_G	+WC_{BPT}	+WC_{CC}	+WR_{FTC}	+GF	+RS_{FD}	ST	BL	TD	TDT_E
10-types	5.3	34.8	36.7	41.6	41.0	43.3	40.9	40.0	36.2	35.1	46.4	47.3
company	0.0	30.0	34.5	33.3	35.2	33.3	32.0	33.3	27.7	26.2	42.1	46.2
facility	0.0	12.4	9.6	20.8	18.6	17.9	14.5	16.7	30.4	19.2	37.5	34.8
geo-loc	5.2	47.2	48.1	53.8	54.4	55.9	56.7	56.1	49.7	48.4	70.1	71.0
movie	8.0	7.4	6.5	8.3	7.7	9.5	23.5	28.6	8.3	0.0	0.0	0.0
musicartist	0.0	6.6	8.5	9.1	9.5	12.7	6.5	14.7	0.0	0.0	7.6	5.8
other	5.8	18.6	18.7	22.5	20.9	26.6	22.1	17.7	24.2	27.7	31.7	32.4
person	11.4	55.1	58.5	63.4	63.8	64.8	65.0	60.2	53.4	50.2	51.3	52.2
product	2.9	12.7	20.0	16.7	18.2	15.4	10.8	11.9	9.0	11.9	10.0	9.3
Sportsteam	0.0	12.9	27.9	30.5	29.0	28.1	27.7	25.4	12.8	13.1	31.3	32.0
tvshow	0.0	0.0	0.0	16.7	16.7	16.7	18.2	13.3	0.0	14.3	5.7	5.7
No-types	13.1	48.3	52.5	56.7	56.4	57.4	53.7	52.9	50.5	51.7	57.3	59.0

3.3 Deep multi task multi dataset learning

Multi task learning [Caruana, 2006, Caruana, 2012] deals with the aspect of training a machine learning algorithm on related tasks so as to enable it to learn robust patterns. With the recent advances in deep learning [LeCun et al., 2015, Bengio, 2009, Schmidhuber, 2015], and the usage of back-propagation to train end to end modular neural networks, it is possible to easily build models which can be trained on multiple tasks and across multiple datasets, while making them learn hierarchical patterns of the data which are relevant for each task. In this chapter our focus will be on building one such multi-task learning model which can be applied to information extraction tasks for social media data. Specifically, we will be considering the tasks of sentiment prediction, named entity recognition and linking, part of speech tagging, rumor classification, twitter author profiling, and tweet and user geo-location prediction.

Why multi task learning for social media? Annotated data for social media IE tasks is very sparse and usually doesn't follow consistent annotation practices. This results in large amounts of noisily annotated dataset which cannot be made to inter-operate with each other. Using multi-task learning we can overcome this limitation and allow the algorithm the identify common patterns across tasks and datasets.

3.3.1 Deep Learning for Information Extraction

Deep learning allows for learning hierarchical representation of the input and its mapping with the output space in an end to end fashion using back-propagation algorithm. We wish to utilize this features for multi task learning. Specifically, internal representations of the model can be utilized for multiple tasks and intermediate representations between tasks can also be combined with each other. Figure 3.6 shows an example of this approach. Where the input is encoded into an internal representation using the encoder. The decoder then decodes the representation and maps it to the output (in this case named entity tags).

3.3.2 Deep multi-task learning

The architecture shown in figure 3.6 can be extended for multi task learning. The core idea here is we need to ensure that the model learns across multiple tasks. This can be achieved by sharing an encoder layer for multiple tasks and having a separate decoder for each task. In the NLP and IE domain this kind of model architecture was popularized in [Collbert et al., 2011]. Recent advancements in improving training of deep neural networks for NLP has shown multiple applications of this approach in the NLP domain [Peng and Dredze, 2015]. However, there doesn't exist proper evaluation of these techniques for social media data.

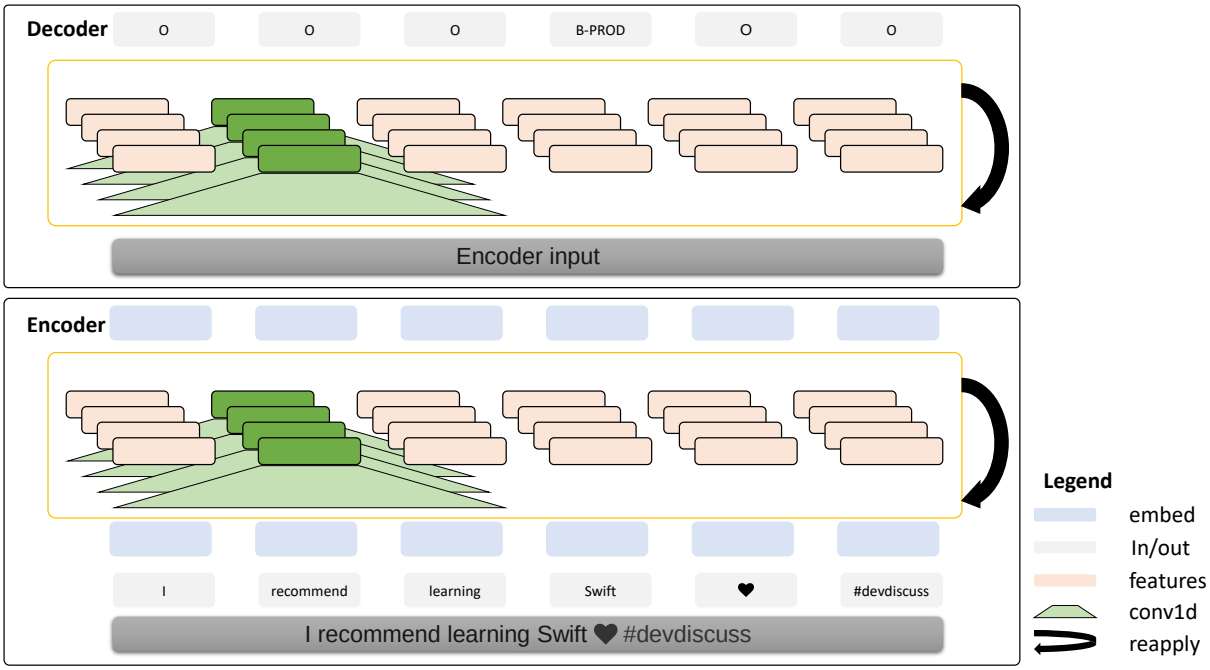


Figure 3.6: Encoder decoder framework for learning hierarchical representations of the data. The example presented is for the named entity recognition task.

3.3.3 Background

Inspired from the recent applications of deep neural networks [Rei, 2017, Balikas et al., 2017, Søgaard and Goldberg, 2016, Liu et al., 2017] for multi task learning in IE and NLP domain, we assess its use on the low resource social media datasets.

3.3.4 Dataset

This research work is possible because of the availability of multiple corpora extracted from social media websites (primarily twitter) as well as the significant improvement shown by deep learning systems on supervised and semi-supervised learning tasks. The following are a list of datasets we plan to use for the various information extraction tasks:

To the best of our knowledge, this is the first instance of combined evaluation of a system of multiple information extraction tasks which include single label as well as sequence label prediction tasks.

3.3.5 Model

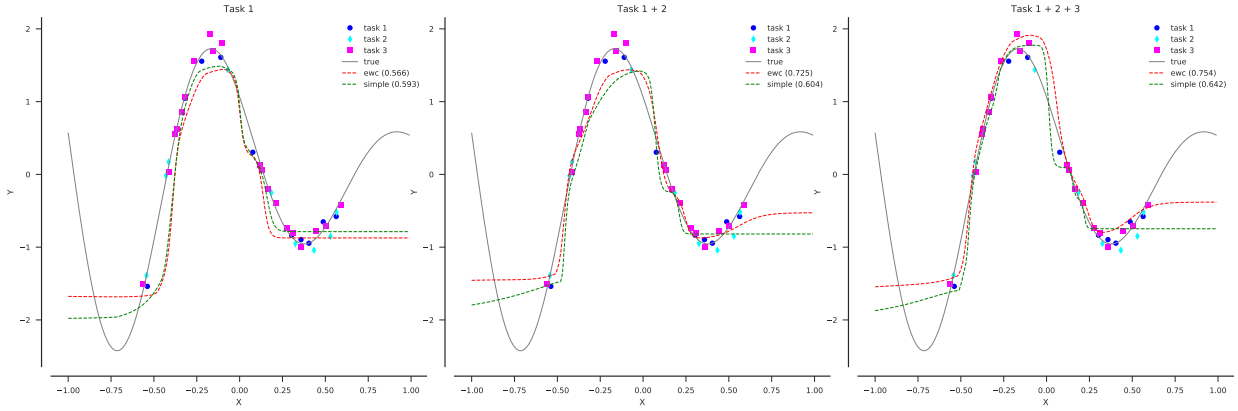
We use the recently proposed elastic weight consolidation model as proposed in [Kirkpatrick et al., 2017] as a major component of our model.

Table 3.3: List of datasets used for social media analysis

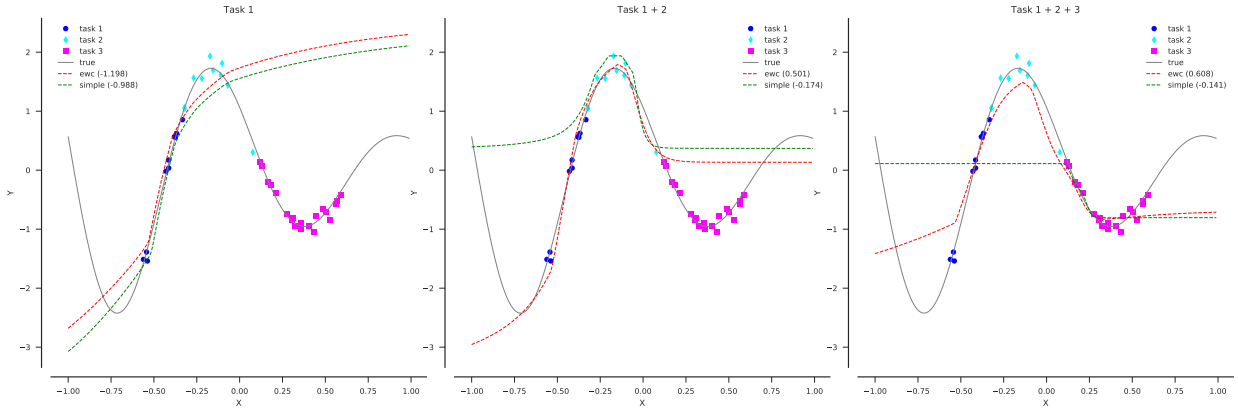
Task	**Data Citations**
Twitter POS	[Derczynski et al., 2013, Owoputi et al., 2012, Owoputi et al., 2013]
Twitter NERC	[Strauss et al., 2016, Baldwin et al., 2015, Ritter et al., 2011]
Twitter NELD	[Dredze et al., 2016, Derczynski et al., 2015, Locke, 2009, Habib and Keulen, 2012, Rizzo et al., 2016]
Twitter Sentiment	[Nakov et al., 2016b, Nakov et al., 2016a, Mozetič et al., 2016, Mishra et al., 2014]
Twitter rumor response classification	[Zubiaga et al., 2016b, Zubiaga et al., 2016a]
Twitter Geolocation Prediction	[Han et al., 2016]
Twitter author profiling	[Rosso et al., 2016]

Table 3.4: Description of datasets for named entity recognition on Tweets

datakey	datatype	# Sequences	# Tokens	# Labels	Labels
Finin	test	2975	51056	3	LOC, PER, ORG
	train	10000	172188	3	LOC, PER, ORG
Hege	test	1545	20664	3	LOC, ORG, PER
MSM_2013		1450	29089	4	LOC, MISC, ORG, PER
	train	2815	51521	7	MISC, PER, ORG, MISRC, LOC, ORG, ORG
Ritter	test	2394	46469	3	LOC, PER, ORG
WNUT_2016	dev	1000	16261	10	PRODUCT, MUSICARTIST, COMPANY, PERSON, FACILITY, OTHER, MOVIE, SPORTSTEAM, TVSHOW, GEO-LOC
	test	3850	61908	10	FACILITY, PRODUCT, MUSICARTIST, COMPANY, PERSON, SPORTSTEAM, OTHER, MOVIE, TVSHOW, GEO-LOC
	train	2394	46469	10	PRODUCT, SPORTSTEAM, COMPANY, PERSON, FACILITY, OTHER, MOVIE, TVSHOW, GEO-LOC, MUSICARTIST
WNUT_2017	dev	1009	15733	6	CREATIVE-WORK, PERSON, CORPORATION, PRODUCT, LOCATION, GROUP
	test	1287	23394	6	CREATIVE-WORK, PERSON, CORPORATION, PRODUCT, LOCATION, GROUP
	train	1000	62730	6	CREATIVE-WORK, PERSON, CORPORATION, PRODUCT, LOCATION, GROUP



(a) i.i.d. samples observed



(b) sequentially observed samples

Figure 3.7: Continual learning on data with common distribution but differently observed samples, (ewc) denotes model trained using elastic weight consolidation, (simple) means models fine-tuned on each newly observed sample. Number in legend depict R-squared value of the model.

Chapter 4

Applications and presentation of extracted information

4.1 SCTG: Social Communications Temporal Graph A novel approach to visualize temporal communication graphs from social data

In this section we present an approach to visualize social conversations using information extraction markers such as sentiment, users, key dates, etc.

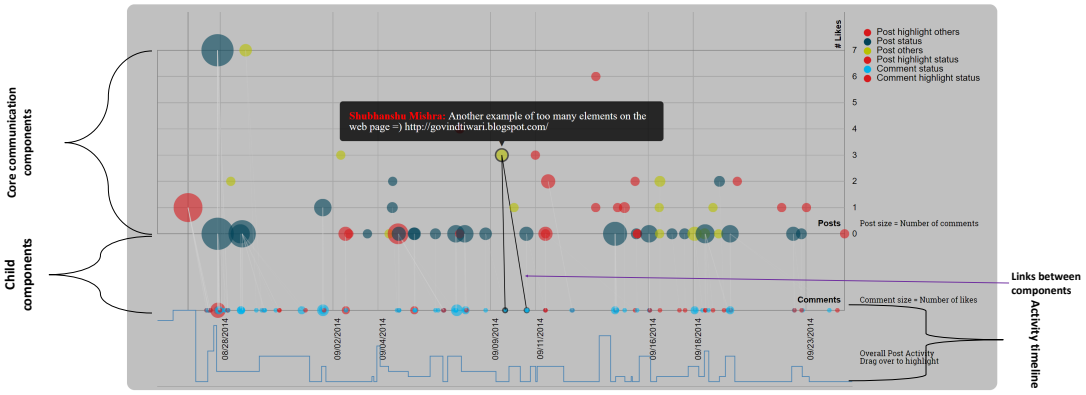
Communication on social channels such as social media websites, email, forums, and groups; follows an inherent temporal network structure. Herein, each communication e.g. a post, occurs at a specific point in time, which can be extracted from its metadata. Furthermore, each communication is also linked to a creator e.g. a user, organization, topic, or another communicationwhich created the communication. Finally, the communication items can be tagged with additional numeric metadata which can be used to score some attributes about the communication e.g. number of comments, retweets, or shares. Existing timeline or network visualizations are not able to do justice to the temporal network structure of such communications.

We present Social Communications Temporal Graph (SCTG) [Mishra, 2017] which is a framework for visualizing such trends. An example of such kind of visualization is presented in 4.1, which shows how the temporal evolution of conversations on a Facebook course group can be visualized. SCTG is a web based visualization which builds on the visualization principles of overview, zoom, filter, details-on-demand, relate, history, and extracts [Shneiderman,]. SCTG is aimed at highlighting the temporal communication nature of social communication channels while allowing various meta-data attributes to be shown alongside.

4.1.1 Background

Our work is mostly inspired from the overview, zoom, filter, details-on-demand, relate, history, and extracts – theory presented in [Shneiderman,]. However, there exists a vast literature on visualizing dynamic and temporal graphs. One particular instance of this is the TimeArcs [Dang et al., 2016] interface which is most

Figure 4.1: Visualization of conversation growth in a Facebook course group



similar to our visualization. It provides an interface for visualizing dynamic network of entities. However, the network consist of no-meta data information about the entities. The visualization is aimed at exploring the evolving relationship between entities e.g. actors in the IMDB network, of named entities in blog corpora. [Beck et al., 2017] provide a comprehensive overview of dynamic graph visualization techniques in existing literature. Our work is likely to fall under the **Timeline** → **Nodelink** → **Integrated layout** as per the taxonomy presented in [Beck et al., 2017]. Our work is very similar to the work of [Reitz, 2010], which uses node scaling and coloring properties to visualize the ego network of an author. However, our approach differs by allowing the user to visualize networks of each entity in the complete network, using the hover option. The work of [Shi et al., 2015] is also related. However, they use a common timeline distribution and connect the network on top of it, they do not account for links between different kind of entities. Finally, most of the aforementioned work is focused on presenting the visualization of ego-networks with examples from scholarly data. Our work enables extending these approaches to generic social network data.

Similarly, prior work on discourse visualization have utilized networks for presenting the discourse structure. For example, the NEREx framework [El-Assady et al., 2017], allows for a comprehensive visualization of debate transcripts using named entity relation graphs. They provide multiple views to explore the data. Which allows for close as well as distant reading of the corpus.

The utility of this visualization can be evaluated by studying the correlation between the values of various visualization components and a ground truth utility e.g. prominent node sizes and their status in the social network.

4.1.2 Components

- **Core communication components**: This can be a user in a feed or a specific post

- **Child components**: This can be associated posts by a user or comments to a post
- **Component links**: Core communication is liked to its children
- **Activity timeline**: This quantifies the temporal activity measurement
- **Tool tips**: They provide additional data about each component
- **Component heights, scaling, and color**: Visualize additional metadata.

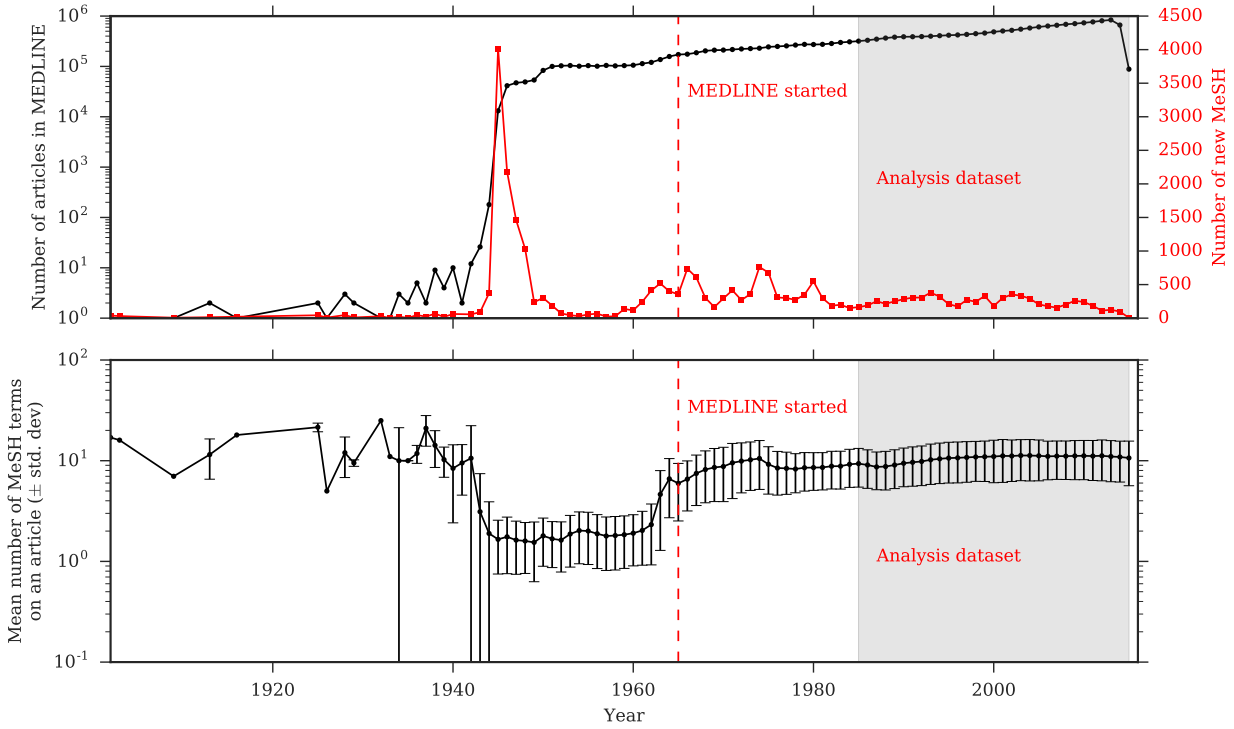
4.1.3 Applications

- **Facebook group data**: Each post in the group feed is a core component, each comment is its children. Posts colored based on content type (e.g. links, text, videos, etc.), scaling based on number of likes.
- **Twitter data with sentiment**: Each user is a core component, each tweet are its children. Tweets can be colored based on sentiment labels, scaled based on retweet counts, and users scaled based on number of followers [see demo]
- **Email data**: Each email is core component, replies are children. Post colored based on email folder, scaling based on number of participants.

4.2 Quantifying Conceptual Novelty in the Biomedical Literature

We introduce several measures of novelty for a scientific article in MEDLINE based on the temporal profiles of its assigned Medical Subject Headings (MeSH). First, temporal profiles for all MeSH terms (and pairs of MeSH terms) were characterized empirically and modelled as logistic growth curves. Second, a paper's novelty is captured by its youngest MeSH (and pairs of MeSH) as measured in years and volume of prior work. Across all papers in MEDLINE published since 1985, we find that individual concept novelty is rare (2.7% of papers have a MeSH 3 years old; 1.0% have a MeSH 20 papers old), while combinatorial novelty is the norm (68% have a pair of MeSH 3 years old; 90% have a pair of MeSH 10 papers old). Furthermore, these novelty measures exhibit complex correlations with article impact (as measured by citations received) and authors' professional age.

Figure 4.2: TOP: Growth of MEDLINE across the years and number of new MeSH terms added each year. BOTTOM: Mean number of MeSH terms used to index articles in MEDLINE across years. MEDLINE was started in 1965 and a lot of noise with regard to many MeSH terms being wrongly spelled or older articles being indexed by too few MeSH terms is present in our corpus around those years. The data after 1985 (shaded grey and marked as Analysis data) has a stable growth and is used for most of the results presented in the Results section.



4.2.1 Data

For generating the novelty scores we consider 22.3 million the articles published in MEDLINE between 1902 and 2015. Our study uses 27,249 MeSH terms as a basis for identifying the concepts of a MEDLINE article. (We use the 2015 MeSH tree.) From Figure 4.2, the rapid growth of MEDLINE after 1945 is quite evident. We also observed that the number of MeSH terms first indexed in a year saw a sharp spike in 1945, and after 1985 this trend has been stable. Similarly, the mean number of MeSH terms in an article has a steady trend of an average of 2 MeSH terms per year since 1985.

4.2.2 Analysis

Figure 4.3 shows the profile of the HIV MeSH term in our dataset. We can observe that the data for HIV fits perfectly with our model and we observe the four distinct phases of growth of the MeSH term. Specifically, 1986 is the year the term enters a Decelerated growth phase and soon after that it enters a Constant growth phase. The observations are interesting because AIDS was first clinically discovered in 1981 and HIV was

discovered in 1983 under two different names LAV and HLTV-III, which matches clearly the accelerated growth in research on this topic. The terms were renamed HIV in 1986. We found that there were no articles mentioning HIV directly before 1986. This also proves why our method is robust in identifying the initial phases of a concept by using the exploded MeSH tree, merging name variants and fixing common spelling issues.

Figure 4.4 shows the relation between the mean impact scores versus the various novelty scores of articles aggregated in time windows of 5 years. The figure depicts a positive correlation between more novel articles and higher impact scores. However, the trends are not consistent across the years and we observed that the articles on novel individual concepts, which were published in more recent years, have lower impact scores. No such trend was visible in articles on novel pairs of concepts in the same years. This might suggest that articles which introduce completely new topics take some time to achieve their potential impact as has been discussed in an earlier literature, however articles which are among the first few to merge existing topics require less time to achieve their potential impact. A possible reason for the observation of this trend might be that there is slow adoption of research in new concepts in the biomedical community, resulting in a low impact of articles published on these concepts in their earlier years, but as the concepts age, a larger number of papers refer to these concepts leading to higher impact later on. We plan to further investigate this effect of novelty on rate of gaining impact in our future analysis.

Figure 4.3: Temporal profile of MeSH term HIV including describing the empirical as well as predicted trends using our model. The figure describes the rapid growth in publications on HIV around 1985 marking a four year period of accelerated growth followed by a three year period of decelerated growth leading into a final phase of constant growth. The model was fitted on the normed count and the predicted values were rescaled to predict the actual number of articles on the concept.

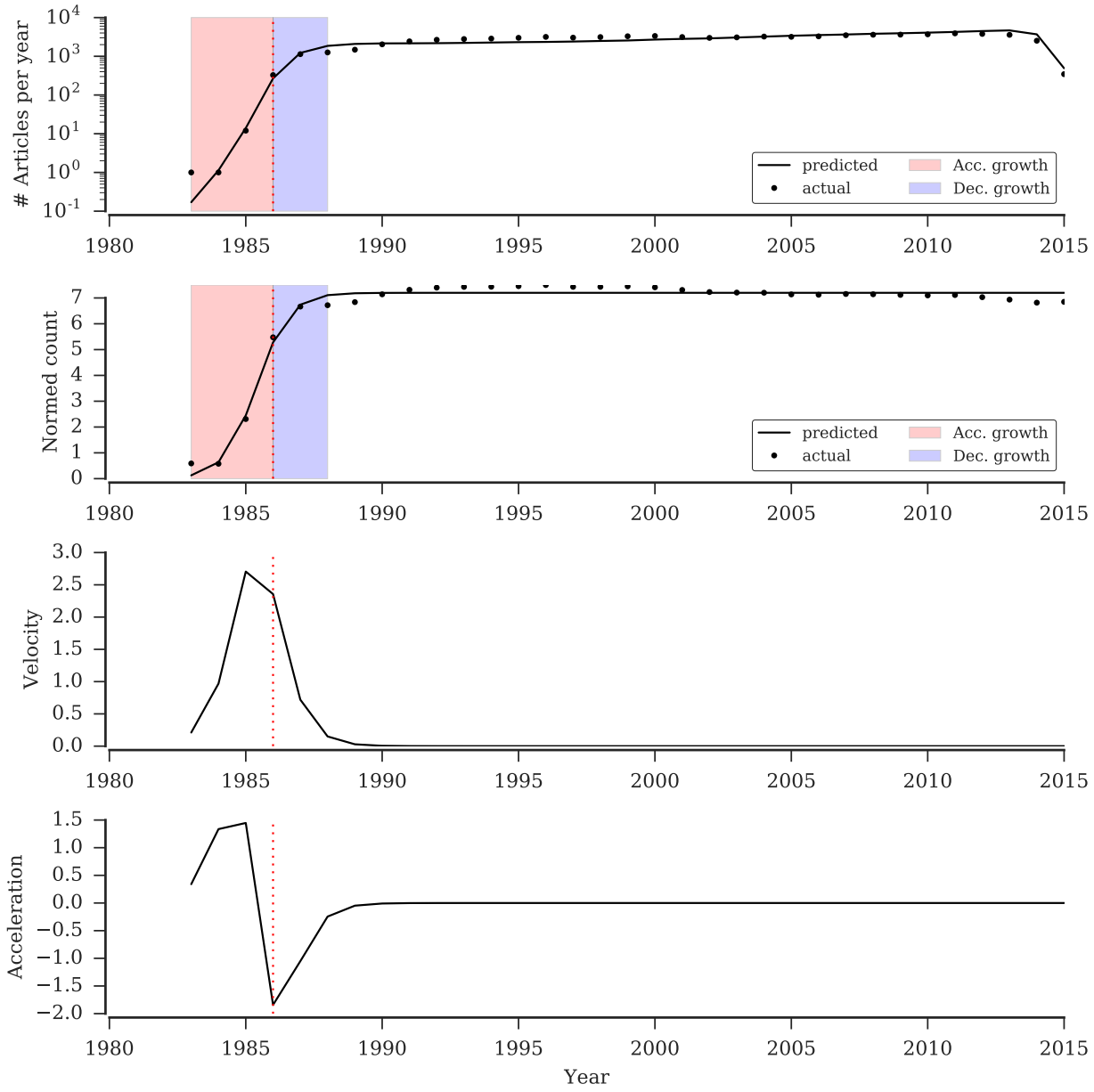
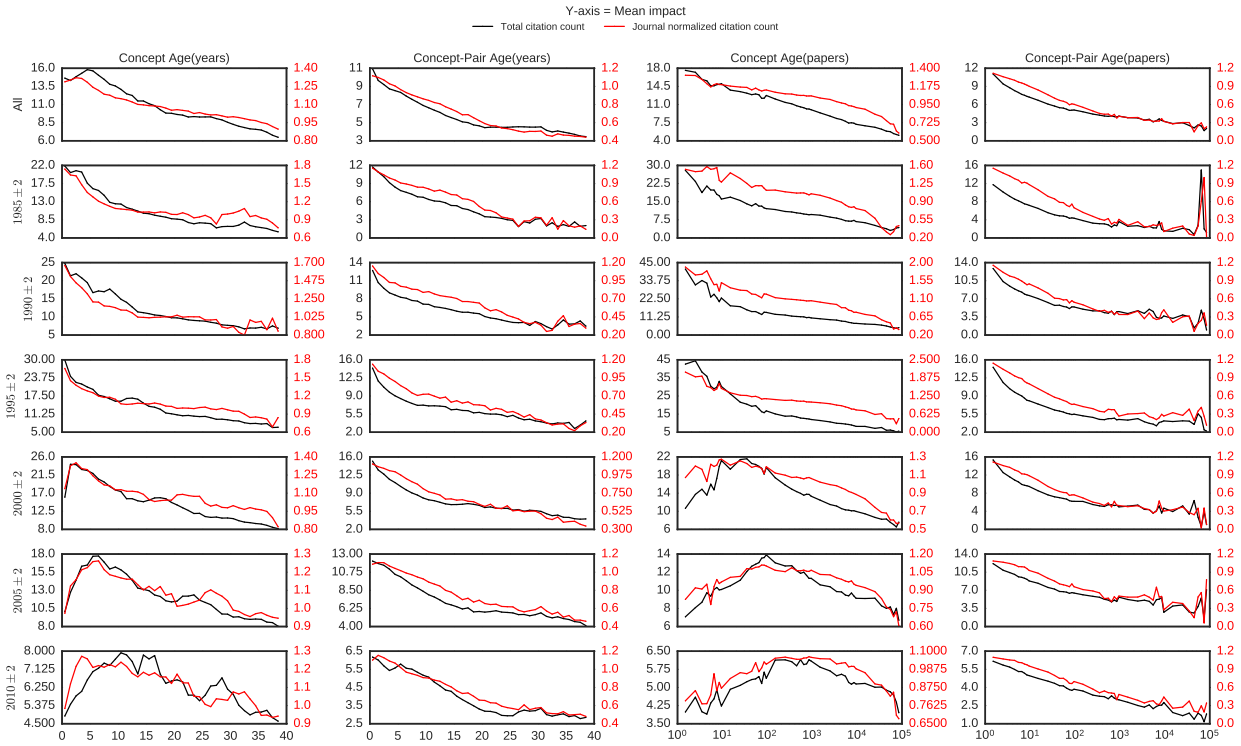


Figure 4.4: Novelty scores correlated with mean impact as measured using total and journal normalized citation count. Y axis in each sub plot represents the mean impact (black total cites, red journal normalized). X axis in each subplot column measures the concept age as denoted in the column title.



Chapter 5

Conclusion

5.1 Limitations

Our work is limited in the following directions:

- Utilizing the DSTD representation is computationally challenging because it breaks the independent identically distributed (iid) assumption of many IE systems used for modeling the data.
- We only present example from sentiment analysis and named entity recognition, which are usually NLP tasks.
- SCTG framework requires evaluation of its effectiveness as a visualization medium.

This work proposes information extraction (IE) systems for social media and scholarly data, which can be identified as instances of digital social trace data (DSTD). DSTD comprise of traces of digital activity generated as part of the interactions. In this work, a DSTD is a representation of social interactions, which includes temporal dependence as well as node and edge level meta-data. This representation allows us to incorporate better contextual information in IE systems. Furthermore, utilizing the DST representation can provide a more holistic view of the underlying social processes.

This proposal focuses on answering the following questions in the domain of IE for DSTD: (RQ1) what information to extract to generate a more interpretable representation of the data? (RQ2) how to extract the information efficiently? and (RQ3) how can the extracted information be presented and utilized? We consider opinion extraction, and named entity recognition as major tasks for social media data.

RQ1 is analyzed using a case of opinion extraction from social media and emphasizes the importance of using domain specific actionable labels such as support versus non-support and enthusiasm versus passivity, as opposed to the traditional positive/negative/neutral labels. Furthermore, emphasis is placed on aggregating opinions at user level, instead of post level, as it is the primary measurement of interest. User level aggregation is further studied using meta-data correlation with human annotated opinion of tweets. Similarly, in the scholarly publishing domain, emphasis is laid on quantifying concept level novelty and expertise

of articles and its authors.

RQ2 is focused on identifying applications of advanced machine learning and parallel processing techniques for improving the efficiency of IE systems on DSTD. In particular, active human-in-the-loop learning, semi-supervised learning, and multi-task learning, are explored. This approach is possible, owing to the recent advancement in training end-to-end model architectures based on deep neural networks. Experiments are conducted to show improvement in sentiment analysis and named entity recognition (NER) tasks for multiple benchmark datasets based on tweets. Parallel processing based algorithms are developed for efficiently computing novelty, and expertise on large scale scholarly datasets.

RQ3 aims at highlighting application and presentation of information extracted using the systems described above. In particular, the thesis presents a novel visualization framework called Social Communications Temporal Graphs (SCTG), which allows its user to visualize three core features of DSTD–temporal dynamics, social connections, and meta-data information – in a single panel. SCTG framework is demonstrated using examples of visualizing Facebook groups and sentiment annotated Twitter corpus. Similarly, scholarly data visualization is presented using platforms for visualizing conceptual novelty, expertise, and temporal author profiles.

Lastly, the thesis proposes open source sharing of an easy to use social media information extraction system, along with other visualization components, thereby enabling social science researchers to effortlessly leverage our methodological and technical contributions in their own research.

Chapter 6

Thesis timeline

The thesis timeline is presented in Fig. 6.1. I wish to defend my thesis by the end of **April, 2018**. In addition, to the timeline, I plan to submit the work done as part of the thesis to the following venues:

- **NAACL-HLT** - December 2017
- **ICWSM** - January 2018
- **IJCAI** - January 2018
- **ACL** - February 2018
- **COLING** - March 2018

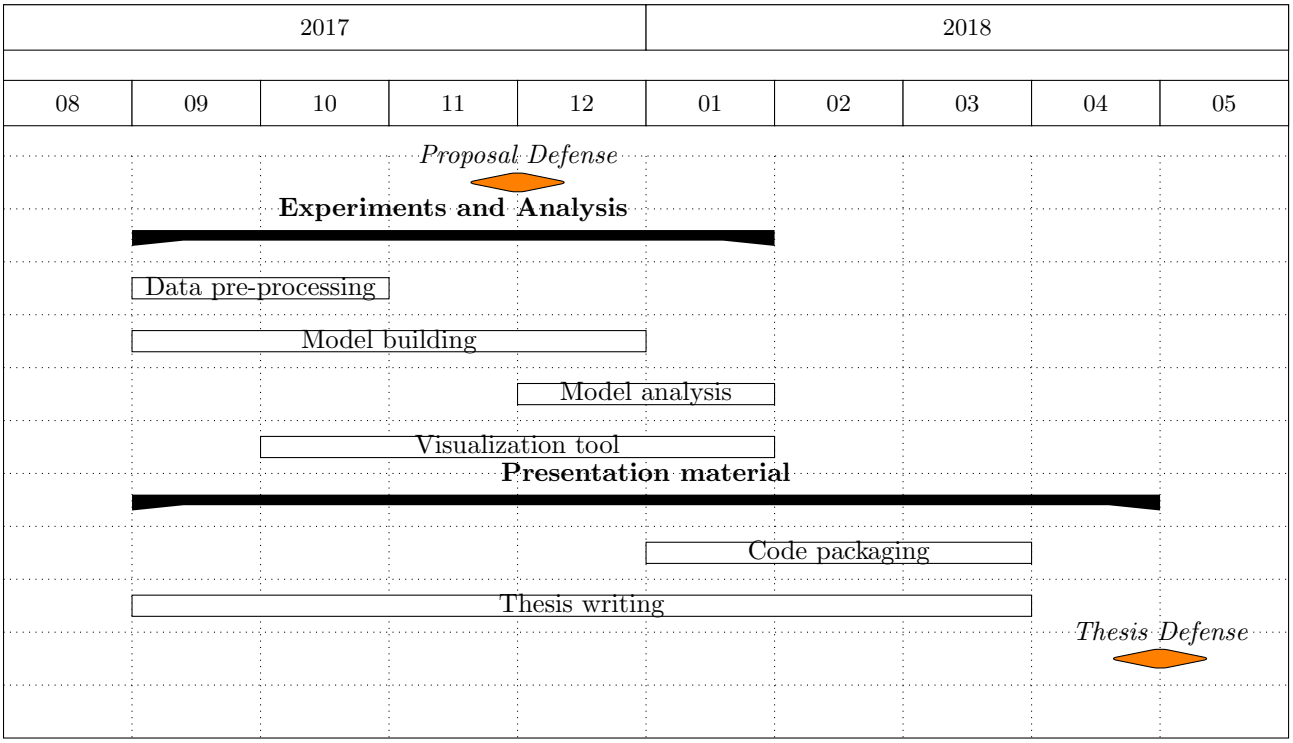


Figure 6.1: Thesis timeline

References

[Aue and Gamon, 2005] Aue, A. and Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 3(3):16–18.

[Baldwin et al., 2015] Baldwin, T., de Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Balikas et al., 2017] Balikas, G., Moura, S., and Amini, M.-R. (2017). Multitask Learning for Fine-Grained Twitter Sentiment Analysis. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, pages 1005–1008, New York, New York, USA. ACM Press.

[Beck et al., 2017] Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2017). A Taxonomy and Survey of Dynamic Graph Visualization. *Computer Graphics Forum*, 36(1):133–159.

[Bengio, 2009] Bengio, Y. (2009). *Learning Deep Architectures for AI*, volume 2. Now Publishers Inc.

[Bollen et al., 2011] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[Bottou, 1991] Bottou, L. (1991). Stochastic Gradient Learning in Neural Networks. In *Proceedings of Neuro-N1imes 91*, Nimes, France.

[Bottou, 2010] Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, Heidelberg.

[Caruana, 2006] Caruana, R. (2006). An empirical comparison of supervised learning algorithms. *conference on Machine learning*, C:161–168.

[Caruana, 2012] Caruana, R. (2012). A Dozen Tricks with Multitask Learning. 0(1998):163–189.

[Chapelle et al., 2006] Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-supervised learning*. MIT Press.

[Collbert et al., 2011] Collbert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal ofMachine Learning Research*, 12:2493–2537.

[Dang et al., 2016] Dang, T. N., Pendar, N., and Forbes, A. G. (2016). TimeArcs: Visualizing Fluctuations in Dynamic Networks. *Computer Graphics Forum*, 35(3):61–69.

[Derczynski et al., 2016] Derczynski, L., Bontcheva, K., and Roberts, I. (2016). Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179.

[Derczynski et al., 2015] Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

[Derczynski et al., 2013] Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206.

[Diesner and Chin, 2015] Diesner, J. and Chin, C.-L. (2015). Usable ethics: practical considerations for responsibly conducting research with social trace data. *Proceedings of Beyond IRBs: Ethical Review Processes for Big Data Research.*

[Dredze et al., 2016] Dredze, M., Andrews, N., and DeYoung, J. (2016). Twitter at the Grammys: A Social Media Corpus for Entity Linking and Disambiguation. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 20–25, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Eisenstein, 2013] Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

[El-Assady et al., 2017] El-Assady, M., Sevastjanova, R., Gipp, B., Keim, D., and Collins, C. (2017). NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. *Computer Graphics Forum*, 36(3):213–225.

[Felbo et al., 2017] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1616–1626.

[Go et al., 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Processing*, 150(12):1–6.

[Habib and Keulen, 2012] Habib, M. B. and Keulen, M. v. (2012). Unsupervised improvement of named entity extraction in short informal context using disambiguation clues. In *Workshop on Semantic Web and Information Extraction, SWAIE 2012*, pages 1–10. CEUR-WS.org.

[Han et al., 2016] Han, B., Rahimi, A., Derczynski, L., and Baldwin, T. (2016). Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217.

[Holme and Saramäki, 2012] Holme, P. and Saramäki, J. (2012). Temporal networks.

[Howison et al., 2000] Howison, J., Wiggins, A., and Crowston, K. (2000). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, 12(12).

[Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526.

[Kosinski et al., 2015] Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543–556.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 591, New York, New York, USA. ACM Press.

[Lazer et al., 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Social science. Computational social science. *Science (New York, N.Y.)*, 323(5915):721–3.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Liu et al., 2017] Liu, P., Qiu, X., and Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Locke, 2009] Locke, B. W. (2009). *Named entity recognition: Adapting to microblogging.* PhD thesis, University of Colorado.

[Maynard et al., 2012] Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. *LREC 2012 Workshop @NLP can u tag #usergeneratedcontent*, page 8.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Nips*, pages 1–9.

[Miller, 2011] Miller, G. (2011). Sociology. Social scientists wade into the tweet stream. *Science (New York, N.Y.)*, 333(6051):1814–5.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

[Mishra, 2017] Mishra, S. (2017). SCTG: Social Communications Temporal Graph A novel approach to visualize temporal communication graphs from social data.

[Mishra et al., 2014] Mishra, S., Agarwal, S., Guo, J., Phelps, K., Picco, J., and Diesner, J. (2014). Enthusiasm and support. In *Proceedings of the 2014 ACM conference on Web science - WebSci '14*, pages 261–262, New York, New York, USA. ACM Press.

[Mishra et al., 2015] Mishra, S., Diesner, J., Byrne, J., and Surbeck, E. (2015). Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pages 323–325, New York, New York, USA. ACM Press.

[Mishra and Torvik, 2016] Mishra, S. and Torvik, V. I. (2016). Quantifying Conceptual Novelty in the Biomedical Literature. *D-Lib Magazine*, 22(9/10).

[Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, 2(SemEval):321–327.

[Mozetič et al., 2016] Mozetič, I., Grčar, M., Smailović, J., Alani, H., Mozetič, I., and Scala, A. (2016). Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PLOS ONE*, 11(5):e0155036.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Nakov et al., 2016a] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016a). SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Nakov et al., 2016b] Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., and Zhu, X. (2016b). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

[Nivre et al., 2016] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

[Owoputi et al., 2012] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N. (2012). Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. *Cmu-Ml-12-107*.

[Owoputi et al., 2013] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. a. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *Proceedings of NAACL-HLT 2013*, (June):380–390.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(12):1–135.

[Peng and Dredze, 2015] Peng, N. and Dredze, M. (2015). Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. In *Emperical Methods for Natural Language Processing*, number September.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

[Plank et al., 2014] Plank, B., Hovy, D., McDonald, R., and Søgaard, A. (2014). Adapting taggers to Twitter with not-so-distant supervision. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1783–1792.

[Pontiki et al., 2015] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Rei, 2017] Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Reitz, 2010] Reitz, F. (2010). A Framework for an Ego-centered and Time-aware Visualization of Relations in Arbitrary Data Repositories.

[Ritter et al., 2011] Ritter, A., Clark, S., and Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of Emperical Methods for Natural Langauge Processing*, pages 1524–1534.

[Rizzo et al., 2016] Rizzo, G., Erp, M. v., Plu, J., and Troncy, R. (2016). Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. In *Workshop on Making Sense of Microposts (#Microposts2016)*, Montréal.

[Rosso et al., 2016] Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. (2016). Overview of PAN16: New challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9822, pages 332–350. Springer, Cham.

[Saif et al., 2013] Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset , the STS-Gold Conference Item. In *Proceedings of the 1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM'13)*, pages 9–21, Turin, Italy.

[Sarawagi, 2008] Sarawagi, S. (2008). Information extraction. *Foundation and Trends in Databases*, 1(3):261–377.

[Schmidhuber, 2015] Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[Settles, 2009] Settles, B. (2009). Active Learning Literature Survey. Technical report, University of Wisconsin–Madison.

[Settles and Craven, 2008] Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks.

[Settles et al., 2008] Settles, B., Craven, M., and Friedland, L. (2008). Active Learning with Real Annotation Costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*.

[Shi et al., 2015] Shi, L., Wang, C., Wen, Z., Qu, H., Lin, C., and Liao, Q. (2015). 1.5D Egocentric Dynamic Network Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(5):624–637.

[Shneiderman,] Shneiderman, B. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Comput. Soc. Press.

[Socher et al., 2013] Socher, R., Chen, D., Manning, C. D., and Ng, A. (2013). Reasoning With Neural Tensor Networks for Knowledge Base Completion.

[Søgaard and Goldberg, 2016] Søgaard, A. and Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Strauss et al., 2016] Strauss, B., Toma, B., Ritter, A., Marneffe, M.-C. d., and Xu, W. (2016). Results of the WNUT16 Named Entity Recognition Shared Task. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

[Sun and Han, 2012] Sun, Y. and Han, J. (2012). Mining Heterogeneous Information Networks: Principles and Methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159.

[Tjong Kim Sang and De Meulder, 2003] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, volume 4, pages 142–147, Morristown, NJ, USA. Association for Computational Linguistics.

[Tumasjan et al., 2010] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *International AAAI Conference on Web and Social Media*.

[Wilson et al., 2012] Wilson, R. E., Gosling, S. D., and Graham, L. T. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*, 7(3):203–220.

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. In *Acl*, volume 7, pages 12–21, Morristown, NJ, USA. Association for Computational Linguistics.

[Zhu, 2008] Zhu, X. (2008). Semi-Supervised Learning Literature Survey. Technical report, Computer Sciences, University of Wisconsin-Madison.

[Zubiaga et al., 2016a] Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016a). PHEME rumour scheme dataset: journalism use case.

[Zubiaga et al., 2016b] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., and Tolmie, P. (2016b). Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE*, 11(3):e0150989.