A Survey on Applications of Natural Language Processing

Qianlin Liang
University of Massachusetts, Amherst
qliang@cs.umass.edu

Abstract

Advances in Natural Language Processing (NLP) as well as the explosive growth of text data have allowed researchers to build state-of-the-art methods to solve a wide variety of natural language tasks, such as text classification, part-of-speech tagging and semantics parsing. In this survey, we describe how these methods can be used to build powerful tools to solve problems in various fields, including politics, medicine, education and finance. We will organize and give an overview of researches in applications of NLP, which encompass exciting research and practical challenges as well as promising business opportunities.

1 Introduction

The development of Natural Language Processing as a scientific discipline generally started in 1950s. One of the objectives of NLP is for computer to achieve human-like comprehension of text and language. While the early NLP work focus on machine translation, with the progress of research and increasing amount of available data, a wide variety of related topics, such as sentiment analysis/classification, part of speech tagging (POS) and dialogue system, were included in the studies. Recently, with deep neutral network-based approach become widespread in NLP, researchers can achieve state-of-the-art results in many natural language tasks. For example, (Devlin et al., 2019) shows that they achieved 86.7% accuracy in MNLI (Williams et al., 2018) entailment classification task.

While researches in these traditional natural language tasks continue to progress, modern phenomena and technologies such as IoT, social media and 5G have leaded to the emergence of many new research opportunities and challenges in specific domains, including politics, medicine, education, internet and finance. In this survey, we aim to provide an overview of where NLP can be applied with high impact in these domains. We identify many problems that require conceptual innovation and can advance the field of NLP, as well as being impactful to the society.

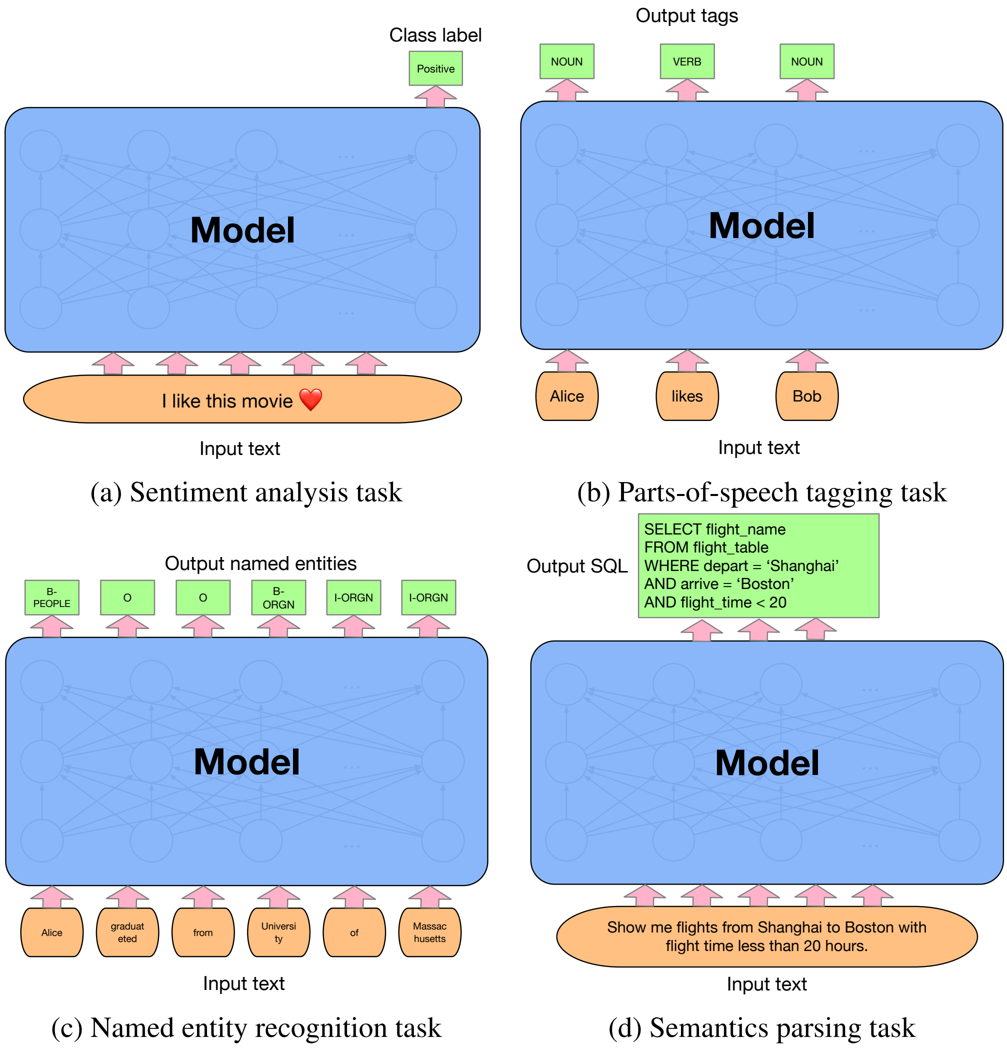
This survey is organized as follow. Firstly, to help readers be familiar with the terminology and concept, section 2 describes general NLP tasks. Then, in section 3, we describe the applications of NLP in the field of internet and cybersecurity. In section 4, we shows applications in medicine. Section 5 shows applications in politics and section 6 shows application in finance. Section 7 and 8 show applications in security and education respectively. Section 9 concludes this survey.

2 Natural Language Tasks

Natural language processing algorithms transmute raw text into any desire structure, such as categorical insight (§2.1), word-level parts-of-speech (§2.2), named entity (§2.3), logic-based representation of meaning (§2.4), translation or conversation (§2.5). Modern NLP algorithms are often developed using *learning algorithms* with labeled *training data*. The labels of the training data have to be well-designed to meet the requirement of the tasks. Table 1 shows a match of application domains and NLP tasks.

2.1 Sentiment Analysis

Sentiment analysis refers to the use of natural language input to extract and identify subjective information. Specifically, the input of a sentiment analysis system is raw text and the output is categorical label in some pre-defined set. However, in contrast to traditional document classification, the labels are defined by sentiment instead of topic. One of examples of sentiment analysis is (Pang



(a) Sentiment analysis task

(b) Parts-of-speech tagging task

(c) Named entity recognition task

(d) Semantics parsing task

Figure 1: Procedures of general Natural Language Processing Tasks.

et al., 2002), which is considered as one of the earliest paper in the field. The authors used machine learning models to extract and convert movie ratings into one of three categories: positive, negative and neutral. Figure 1 (a) illustrates the procedure of sentiment analysis. Sentiment analysis is widely applied to customer feedback analysis such as reviews, online and social media.

2.2 Parts-of-Speech Tagging

Parts-of-speech (POS) is the syntactic role of each word in a sentence. In this task, given raw text, our goal is to label each word with their corresponding POS. For example, given text *Alice likes Bob*. We would like to label *Alice* and *Bob* with `NOUN` and *likes* with `VERB`. Figure 1 (b) shows the procedures of this POS tagging. In contrast to sentiment analysis, which outputs single label, POS tagging outputs a sequence of labels. This is usually done using *sequence labeling* technique in NLP. POS is usually applied as features for downstream tasks such as sentiment analysis and used to check fluency of generated text.

2.3 Named Entity Recognition

Given raw text, the goal of Named Entity Recognition (NER) is to extract named entities mentioned in the text. Figure 1 (c) shows an example of NER. Here, the named entities include *Alice*, people and *University of Massachusetts, Amherst* an organization. This task looks similar to the parts-of-speech tagging - both of them are labeling a sequence. However, they significantly different. Instead of labeling each single word, NER labels a subsequence of words such as *University of Massachusetts, Amherst*. This can be achieved using **BIO notation**. The start word of a named entity is labeled with a prefix of `B-`; The word within a

	Sentiment Anal.	POS Tagging	NER	Semantics Pars.	Dialogue Sys.
Internet					
Hate speech detection	•	•			•
Cybersecurity threats detection	•				
Medicine					
Influenza forecasting	•				
Clinical conversation properties			•		
Politics					
Election outcome prediction	•				
Finance					
Financial event extraction			•		
Dialogue fraud detection					•
Security					
Suicide risk detection	•				
Sexual harassment analysis			•		
Education					
Student answer parsing				•	

Table 1: Application domains match with NLP tasks

named entity is labeled with a prefix of `I-`; The words that are not parts of any named entity are labeled with `O`.

2.4 Semantics Parsing

Semantics parsing refers to the task of converting an unstructured natural language utterance to a structured form, which can be understood by machines. NER can be considered as a case of shallow semantics parsing. That is, NER can extract named entities from the text, however, it cannot describe the logical relation between those entities. Deep semantics parsing tries to extract the logic and convert to formal meaning representation language such as SQL. Figure 1 (d) shows an example of semantics parsing.

2.5 Dialogue System

A dialogue system could include all of the tasks previously described. Given input text, a typical cycle in a dialogue system basically contains the following three phases: 1) convert the text to meaning representation language using semantics parser, which may includes POS tagging and NER; 2) keep the history of dialog and retrieve information from knowledge base; 3) produce output using output generator. POS tagging may be used to check the generated text quality here.

3 Internet

Today, a large amount of text data is generated from online applications. Many social media sites, such as Twitter and Reddit, open their APIs to allow developers and researchers collecting information effectively. This leads to an increasing interest in using social media data to address real world problems.

3.1 Online Hate Speech Detection

Hate speech is a particular form of offensive language that express hate or encourage violence toward a person or group based on some stereotypes such as race, religious and sex. Hate speech is not uncommon on the internet and sometime it could culminate in severe threats to individuals. Thus, detecting and censoring these hate speech becomes a challenge for both site developers and NLP researchers.

One of the key challenges in hate speech detection is lack of general definition of hate speech. There are many research works on hate speech detection, however, most of them employ different definition of hate speech. (Warner and Hirschberg, 2012) hypothesize hate speech often employs well known stereotypes to disparage an individual or group and design to annotate their corpus as anti-semitic, anti-black, anti-asian, anti-woman, anti-muslim, anti-immigrant or other-hate. (Waseem and Hovy, 2016) defines a list of 11 rules to identify hate speech, including uses a sexist or racial slur, attacks a minority etc. However, most of these rules are themself ambiguous. (Davidson et al., 2017) indicates that much previous work wrongly labels offensive language as hate speech and tries to separate hate speech from other instances of offensive language.

Although these work employ different definition of hate speech, they all threat the problem as classification problem. After collecting and annotating the corpus, they build classifier using handcrafted features(Yarowsky, 1994) and parts-of-speech tags. (Waseem and Hovy, 2016) also incorporates geographic and demographic features and found that adding gender information improves their result. However, these information may generally be unavailable. The classification results are often evaluated in terms of precision, recall, accuracy, F-1 score and AUC.

In contrast to simply detecting and blocking hate speech, (Qian et al., 2019) argues that this would not mitigate hate speech and is at odds with the concept of free speech. Thus, beside just detecting hate speech, the authors create generative models to generate intervention for hate speech. Moreover, they argues that most of previous work treat each post as an isolated instance, such as tweet, but hate speech often depends on the context. Thus, they develop a conversational-based dataset for hate speech intervention generation. The authors create Seq2Seq models for this task, including RNN, Variational Auto-Encoder (VAE) and Reinforcement Learning (LR) and evaluate them using BLEU(Papineni et al., 2002), ROUGE(Lin, 2004) and METEOR(Banerjee and Lavie, 2005).

3.2 Cybersecurity Analysis

Breaking cybersecurity events are shared across a range of websites including blogs and social media such as Twitter and Facebook. However, there is a large time delay between the time a vulnerability is first publicly disclosed to when it is formally published to the National Vulnerability Database (NVD). As the rate of discovered vulnerabilities has increased in recent year, the need for efficient identification and prioritization has become more important.

(Zong et al., 2019) studies using NLP techniques and Twitter data to analyze users' opinions about the severity of software vulnerability reported online. The authors treat this task as a sentiment analysis problem. Specifically, given a named entities and tweet $< e, t >$, the goal is the predict the probability the tweet describes a cybersecurity threat toward the entities $p_{threat}(y| < e, t >)$ and the severity $p_{severe}(y| < e, t >)$.

The key challenge for this task annotating the dataset, because judging the severity of cybersecurity threat based on text requires expert experience and unexperienced annotators could result in low inter-rater agreement. The authors solve this problem by using multi-phases annotation. In the first phase, the annotators judge whether a tweet contains cybersecurity threats toward target entity. In the second phase, annotators judge the severity of the threat. Only the tweets pass the first phase are annotated by severity.

After getting the data, the authors try several classifier, including Logistic Regression, Convolutional Neutral Network (CNN) and Word Embeddings and achieve the result of Precision@50 of 0.86, which shows that we can use NLP techniques to predict high severity vulnerabilities effectively.

4 Medicine

We describe two applications of NLP in the field of medicine. The first one is sentiment analysis-based influenza forecasting. The second is extracting entities and relations from clinical conversation using named entities recognition.

4.1 Influenza Forecasting

Accurate disease forecasts are imperative when preparing for influenza epidemic outbreaks. Traditionally, influenza forecasting methods are based on historical Influenza-like Illness (ILI) data from the U.S. Centers for Disease Control and Prevention (CDC). (Lamb et al., 2013) describe an influenza surveillance method based on Twitter data. The authors create classifier to classify if a tweet is related to influenza or not. Then, they monitor the trend of related tweets. The result shows high correlation between predicted trend and the ground true ILI data reported by CDC.

(Paul et al., 2014) incorporate predicted result from (Lamb et al., 2013) with historical ILI data to predict the influenza rate in k weeks.

$$y_{w+k} = \alpha_1 \widetilde{y}_{w-1} + \alpha_2 \widetilde{y}_{w-2} + \alpha_3 \widetilde{y}_{w-3} + \gamma_1 z_w \quad (1)$$

Equation 1 describes the model they use to predict influenza rate. Where y_{w+k} is the influenza rate for week $w + k$, α_i, γ_i are parameters, $\widetilde{y}_i$ is historical influenza rate for week i and z_w is the Tweeter data-based predicted value. The authors make comparison between ILI data-only autoregressive method, Tweeter data-only method, Google Flu Trends (GFT) predicted result, ILI data + Tweeter and ILI data + Tweeter + GFT method. The result shows that ILI data + Tweeter

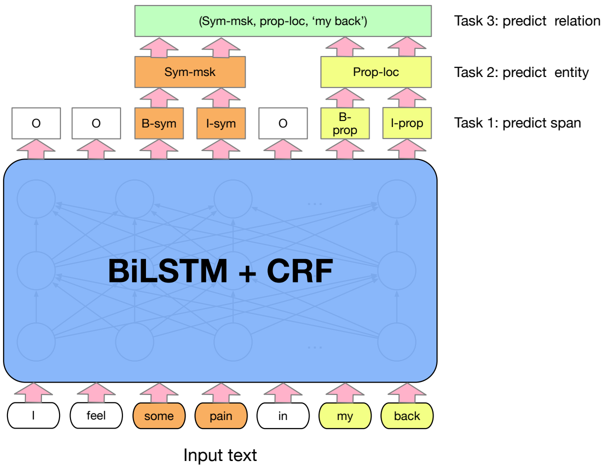


Figure 2: Example of extracting entities, properties and relations from clinical conversations

data method achieve the lowest mean absolute error, which shows that influenza surveillance signal from Twitter significantly improve influenza forecasting.

4.2 Extracting Relations from Clinical Conversation

The widespread adoption of Electronic Health Records has overwhelmed clinical providers. To address the problem, there has been increasing interest to automate aspects of documentation so that clinical providers can spend more time with their patients. One of such approach aims to generate clinical notes directly from doctor-patient conversations.

(Du et al., 2019) propose a method to jointly extract entities, properties and their relations from clinical conversations. The authors define the task as to extract *(symType, propType, propContent)* tuple from the conversations. Where *symType* is symptom, *propType* is the property (e.g. frequency, location) and *propContent* is the content. Figure 2 shows an example of this task.

This task can be converted to a regular Named Entities Recognition task where the entities space is the cross product of the *symType* and *propType*, that is:

$$|entities| = |symType| \cdot |propType|$$

However, this method is infeasible when the symptom type space is too large. The origin paper identify 186 symptom types and 3 property types. In this case, the entity space would be 186×3, which is too large to feed into a CRF layer. Thus, the authors only use this as baseline. To address this problem, the authors break the task into 3 subtasks and train them jointly, as illustrated in figure 2. First, they use a BiLSTM + CRF model detect span of the named entities, just as regular Named Entities Recognition tasks do. Then, the recognized spans are fed to an attribute tagging layer to predict their tags. Finally, they combine the predicted *symType* and *propType* to predict the relations. By using this multi-task learning paradigm, the authors report F-scores of 0.34 and 0.45 for symptom and medication information extractions respectively, which is significantly better than 0.18 and 0.35 for the simple NER baselines and pretty close to human-level recognition (0.51 and 0.52).

5 Politics

In this section, we describe how to apply NLP to extract public opinion for political events and decision making.

5.1 Presidential Approval and Election Outcome Prediction

Knowing the public opinion is crucial for predicting presidential election outcome.Traditionally, polling (i.e. ask random sample of people) is used to measure the population attitudes toward presidential candidates. As the rise of freely available text-based social media data and the advances of NLP, mining public opinion from social media text content could be a faster and less expensively way than polling.

(O'Connor et al., 2010) shows how to employ Twitter data to capture the trend of political opinions for presidential approval and election over the 2008 and 2009 period. The authors use over 1 billion tweets posted over the period and the task is to predict the sentiment of related tweets. One of the key challenges in this task is that the dataset is asymmetric - only 0.1-0.5% of all tweets contain related topics. Thus, it's difficult to build useful models using the entire dataset. To address this problem, the authors filter the dataset using a set of keywords. For example, *obama* for presidential approval. Then, they use subjectivity lexicon to classify each of the filtered tweets as positive or negative - tweets contain positive words are considered as positive and tweets contain negative words are considered as negative. Finally, they compute the ratio between positive tweet count and negative tweet count and use the trend of this sentiment ratio to model the public opinion. The results are compared with sets of polls.

Using only the presence or absence of subjective words in a tweet to predict the sentiment is one the most basic methods and often yield poor classification accuracy. However, the authors argue that the goal of these tasks is to infer the *aggregate* sentiment. A high error rate only implies the classifier is a noisy instrument. With a fairly large number of measurements, the errors will cancel out. Finally, moving average is used to smooth the predicted trend. In evaluation experiment, they found the predicted trend achieve a 0.8 correlation with ground true poll trend for presidential approval rating, which highlight the possibility of using social media text content and NLP as a substitute and supplement for traditional polling.

6 Finance

Financial markets are driven by information.(Mitchell and Mulherin, 1994) shows that there is a significantly positive correlation between the number of daily news announcements and trading activities. (Warner et al., 1988) and (Bonnier and Bruner, 1989) show that a management change news would make significantly positive or negative impact on the company, depending on current performance of the company. This motivates a need of incorporating news factor into financial market trading strategies. Furthermore, as the rise of financial loan applications, automatically detecting identity fraud becomes a crucial challenge. We will see how NLP can be applied to solve these problems.

6.1 Financial Event Extraction

Recently, there is a dramatic rise of volumes of digital financial documents, which contain valuable information for detecting emerging risks and profitable opportunities timely. However, this documents are generally unstructured plain text and cannot be fed directly into a trading strategy system. Thus, extracting and converting key events from the overwhelming financial documents into structure information becomes a challenge for both financial and NLP communities.

Traditionally, Named Entity Recognition is applied to extract structured information from financial text. However, NER works within the sentence scope. (Zheng et al., 2019) indicates two challenges for applying NER to financial documents.The first one is called *arguments-scattering*, which indicates that entities/arguments of one event may scatter across multiple sentences. The second challenge is *multi-event*, that is, a document may contains multiple events. Figure 3 shows an example of document-level event extraction from the origin paper.

To address these problems, the authors propose an end-to-end Doc2EDAG model to map input document to *entity-based directed acyclic graph* (EDAG), where each node is an entity/argument related to the event and edges indicate predefined order of entities for the event. The input is the document representation of size (d_w, H_w, H_s) where d_w is the dimension of word embeddings, H_w is the sentence length and H_s is the number of sentences in a document. The model first performs Named Entities Recognition to identify entity spans in the document. The authors use a transformer, say Transformer-1, with conditional random field (CRF) to do this and yield a word level representation of the same size, (d_w, H_w, H_s). Then, to get the document-level representation, one of the key ideas the authors employ is aggregating word representations over recognized entities and sentences using max pooling, which yields a representation of size (d_w, H_s). After that, they feed it to another transformer, say Transformer-2, to generate the document-level representation. Finally, they use another transformer, say Transformer-3, to generate the DAG, recurrently. To address the *multi-event* problem, the key idea is allowing multi-label classification and predicting next entity based on previous predicted entity sequences in a depth first search style. The authors then evaluate the model on a real-world dataset consisting of Chinese financial announcements and get F-1 score results of 82.3 for single event sets and 67.3 for multi event sets.

When Doc2EDAG describes how to extract events from financial news, (Nuij et al., 2014) shows incorporating news variable really improves automatic stock trading strategies. The authors incorporate news into trading strategies automatically using the following three steps approach: 1) extracting the relevant events, as well as involved entities from the text of the news message, 2) associating an impact score with each of the extracted events and 3) making use of the impact of news events in generate buy or sell signals in trading strategy system.

Structured Event Table of Equity Pledge

Event ID	Pledger	Pledged Shares	Pledgee	Begin Date	End Date	Total Holding Shares	Total Holding Ratio
1	[PER]	[SHARE2]	[ORG]	[DATE1]	[DATE4]	[SHARE5]	[RATIO]
2	[PER]	[SHARE3]	[ORG]	[DATE2]	[DATE4]	[SHARE5]	[RATIO]

ID	Document sentences
5	In **[DATE1]**, **[PER]** pledged his **[SHARE1]** to **[ORG]**.
7	After the company carried out the transferring of the capital accumulation fund to the capital stock, his pledged shares became **[SHARE2]**.
8	In **[DATE2]**, **[PER]** pledged **[SHARE3]** to **[ORG]**, as a supplementary pledge to the above pledged shares.
9	The aforementioned pledged and supplementary pledged shares added up to [SHARE4], and the original repurchase date was **[DATE3]**.
10	In **[DATE3]**, **[PER]** extended the repurchase date to **[DATE4]** for **[SHARE4]** he pledged.
12	As of the date of this announcement, **[PER]** hold **[SHARE5]** of the company, accounting for **[RATIO]** of the total share capital of the company.

Figure 3: Example of document-level event extraction from (Zheng et al., 2019). The bottom table shows sentences in a typical financial document and the top table shows the events extracted from the sentences.

Then, the authors run an experiment to measure the effectiveness of financial news as a trading indicator. They employ a dataset of historical company share prices and a collection of news messages related to these companies from January 1st, 2007 to April 30th, 2007. Next, they compare the financial news trading indicator and other widespread used technical trading indicators including simple moving average (SMA), Bollinger band (BB), the exponential moving average (EMA), the rate of change (RoC) momentum (MOM) and moving average convergence divergence (MACD). The result show that the news indicator is consistently outperformed by SMA and EMA, the best single indicators in this experiment, but only slightly.

Next, the authors add news variable and other technique trading indicator to the trading strategy set to form a trading framework. To optimize the framework, they employ genetic programming (GP), which attempts to improve the fitness of the framework by adding or removing strategies in the strategies set. As a result, the news variable is often included in the optimal trading rules, which indicating the added value of news as a trading strategy.

6.2 Dialogue Based Identity Fraud Detection

Identity fraud refers to using the personal information of other person to commit a crime or deceive other persons or groups. As personal loan becomes widespread, identifying identity fraud becomes a challenge for many financial institutions. (Wang et al., 2019) propose a novel method of detecting identity fraud through dialogue interaction. The key idea is employing *derived* questions from personal information and generate next relevant questions based on the answer. Their experiments show that the dialogue based method achieve an accuracy of 0.884, which is significantly higher than 0.748

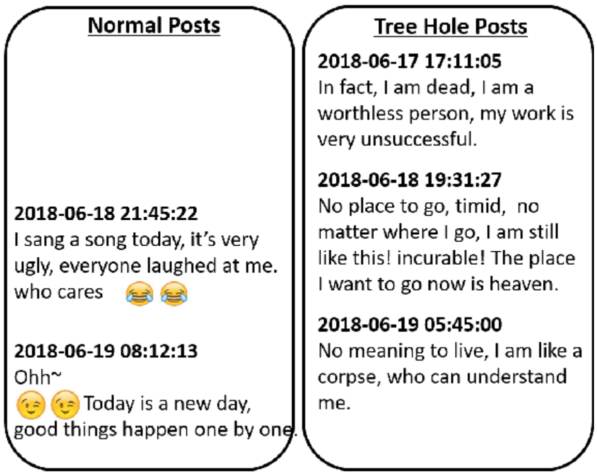


Figure 4: Example of a user's normal vs. anonymous posts from (Cao et al., 2019)

of rule-based method.

7 Security

NLP could be an effective tool for both building security and surveillance systems, and analyzing crime patterns.

7.1 Suicide Ideation Detection

The growing of suicide rate has become one of the major social issues worldwide. Every 40 seconds a person dies by suicide somewhere in the world and many more attempt suicide (Parekh and Phillips, 2014). How to detect suicide ideation and perform intervention timely and effectively has become one of the main concerns to our society.

(Cao et al., 2019) propose a sentiment analysis-based of using social media data to timely detect suicide ideation. However, one of the key challenges is people might hide their inmost thoughts and feelings and express contrarily on social media. Performing a naive sentiment analysis on these

Figure 5: Example of named entities annotation in sexual harassment analysis from (Liu et al., 2019)

posts might be less effective. To address this problem, the authors employ data from tree hole, an anonymous social media application. The authors found that the anonymous posts on tree hole are more self-concerns and express suicidal thoughts more directly. An example from origin paper is show in figure 4.

Using the tree hole data, the authors are able to build suicide-orientated word embeddings and a two-layer LSTM model with attention. The results show that the model with suicide-specific word embeddings achieves 0.91 accuracy.

7.2 Sexual Haraessment Analysis

Recently, there is a growing number of people share their personal stories about sexual harassment online. For example, Safecity, an online forum for people who experienced or witnessed sexual harassment to share their personal experience, has collected over 10,000 stories so far. (Liu et al., 2019) proposes using NLP to extract sexual harassment patterns from these data.

The authors employ CNN-based and BiLSTM-based models to infer the word-level named entities and label for each story jointly. For Name entity Recognition task, there are four entities - harasser, time, location and trigger. An example is shown in figure 5. The story classification task contains 5 subtasks. The model is trained to jointly inference the 1) age of harasser; 2) single/mutiple harasser; 3) type of harasser; 4) type of location; 5) time of the day. The evaluation results show that the jointly trained model outperform the single task models.

Then the authors use the model to profile the sexual harassment stories and describe many insightful findings. For instances, he young harassers often engage in harassment activities as groups and engaged in harassment behavior on the streets. In contrast, adult perpetrators of sexual harassment are more likely to act alone and engage on public transportation. These results can provide valuable information for preventing sexual harassment.

QUESTION	Georgia found one brown mineral and one black mineral. How will she know which one is harder?
REF. ANS.	The harder mineral will leave a scratch on the less hard mineral. If the black mineral is harder, the brown mineral will have a scratch.
STUD. ANS.	The harder will leave a scratch on the other.
LABEL	Correct

Figure 6: Example of a typical student response analysis system from (Dzikovska et al., 2013). Note that student answers skip *brown mineral* and *black mineral* mentioned in the question but the reference answer tend to repeat them.

8 Education

There is an increasing interest in applying NLP in education and communities have emerged to sponsor regular meetings and share tasks (Litman, 2016).

8.1 Student Response Analysis

Student response analysis (SRA is the task of labeling student answers with categories that could help a full dialog system to generate appropriate and effective feedback (Dzikovska et al., 2012). To address this problem, a joint challenge was organized at SemEval-2013 (Dzikovska et al., 2013). Specifically, the task is given a *question*, a known *reference answer* and a 1- or 2- sentence *student answer*, the goal is to label each student answer with one of the five labels - *Correct, Partially correct incomplete, Contradictory, Irrelevant* or *Non domain*. Figure 6 shows an example of this task.

One key challenge is that student answers often skip details mentioned in the question but the reference answer often repeat them. In this case, a semantics predictor might think the student answer does not entail the reference answer and output an incorrect label. One of the approaches is concatenating question and student answer and predict if the concatenated text entail the reference answer. Employing this approach, the best team attended the challenge produced results significantly outperformed the baseline, showing that computational linguistics approaches can contribute to educational tasks effectively.

9 Conclusion

In this survey, we have shown that Natural Language Processing has significant contributions to offer across domain areas. In some domains, NLP methods outperform the traditional methods, and in some domains, NLP can be incorporated to existing systems and significantly improve the performance. Successfully applying NLP to solve real-world problems requires not only state-of-the-arts NLP algorithms, but also domain-specific knowledge. Challenges and opportunities are also highlighted throughout this survey.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Karl-Adam Bonnier and Robert F. Bruner. 1989. An analysis of stock price reaction to management change in distressed firms. *Journal of Accounting and Economics*, 11(1):95 – 106.

Lei Cao, Hui jun Zhang, Ling Feng, ZiHan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *EMNLP/IJCNLP*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2019. Learning to infer entities, properties and their relations from clinical conversations. In *EMNLP/IJCNLP*.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Chris Brew. 2012. Towards effective tutorial feedback for explanation questions: A dataset and baselines. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210, Montréal, Canada. Association for Computational Linguistics.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Diane Litman. 2016. Natural language processing for enhancing teaching and learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI16, page 41704176. AAAI Press.

Yingchi Liu, Quanzhi Li, Xiaozhong Liu, Qiong Zhang, and Luo Si. 2019. Sexual harassment story classification and key information identification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM 19, page 23852388, New York, NY, USA. Association for Computing Machinery.

Mark Mitchell and J Harold Mulherin. 1994. The impact of public information on the stock market. *Journal of Finance*, 49(3):923–50.

W. Nuij, V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak. 2014. An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering*, 26(4):823–835.

Brendan T. O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 02, page 311318, USA. Association for Computational Linguistics.

Amrita Parekh and Michael Phillips. 2014. Preventing suicide: A global imperative.

Michael J. Paul, Mark Dredze, and David A. Broniatowski. 2014. Twitter improves influenza forecasting. In *PLoS currents*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Weikang Wang, Jiajun Zhang, Qian Li, Chengqing Zong, and Zhifei Li. 2019. Are you for real? detecting identity fraud via dialogue interactions. In *EMNLP/IJCNLP*.

Jerold B. Warner, Ross L. Watts, and Karen H. Wruck. 1988. Stock prices and top management changes. *Journal of Financial Economics*, 20:461 – 492. The Distribution of Power Among Corporate Managers, Shareholders, and Directors.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

Shi Zong, Alan Ritter, Graham Mueller, and Evan Wright. 2019. Analyzing the perceived severity of cybersecurity threats reported on social media. In *NAACL-HLT*.