Social media has created a very dynamic and exciting environment for the field of natural language processing. There is an endless stream of data just pouring into the internet via online platforms with every context imaginable. This has created many new challenges and interesting solutions to NLP. In this paper I will be analyzing two chapters from *Natural Language Processing for Social Media* by Anna Atefeh Farzindar and Diana Inkpen [1]. I will be analyzing the contents of chapter 2 “Linguistic Pre-processing of Social Media Texts” and chapter 5 “Data Collection, Annotation, and Evaluation”.

Chapter two focused on Natural Language Processing (NLP) of social media data. NLP has been around a topic of study for over half a century. Moreover, it cannot be a more exciting time for NLP researchers now than ever due to the boom of social media like Facebook, Instagram, and especially Twitter. Social media has given us an infinite amount of ever producing textual data from around the world that is easily accessible. This can give us insights we never could before but of course it comes at the cost of new obstacles for NLP researchers. First issue with NLP is we need to be able to measure how well or bad an NLP method does when applying to the new data of social media. This chapter briefly explains three basic ways. We can either evaluate the performance manually with individual people examine the output of the data. We can automatically measure the performance using scripts and with data that was previously annotated and approved manually. Lastly, by applying tasks for the NLP method and seeing the success rate of the tasks after tweaking the NLP tools. Automatic scripts are shown to be the most convenient and thus the focus in research when developing better NLP methods.

In general machine learning techniques are applied when developing NLP methods. Supervised learning is very popular, along with Naive Bayes and SVM algorithms. However Neural Networks have shown to have the best performance assuming we have a large training data set [2]. This shows the importance of collecting good annotated data in social media which seems to be the most difficult part. Chapter five of *Natural Language Processing for Social Media* [1], is dedicated to the idea of data collection which I summarize in the following section. Additionally, some more unusual types of classification models are shown to have great value like sequence-to-sequence models, in which the input (text) are a sequence (sentence) and a classification is another sequence of labels instead of traditional discrete models. A great example of the usefulness of this model described in this chapter is by [3] in translating text languages. As an input would be a sequence of words and the output would be another sequence of words.

Touching back on the topic of data source, there are many NLP tools available for developer but the problem is that there are not accustomed to the type of data that is living in social media platforms. This data is inconsistent, very informal, grammatically incorrect, and have many variations depending on the demographics of the authors. There are two primary ways of handling this issue that is discussed in this chapter. We can either transform the social media type text to more formal text to conform to the training data sets already stored for our NLP methods (normalizing text) or we can obviously just re-train our NLP methods using raw social media data (but this can take a lot of time and work to acquire training data). To normalize the data can be very difficult as well as this implies coring the errors that are found in the selected social media text. Many researches have proposed efficient ways at normalizing data

from using context around an error to correct it [4] and applying machine learning techniques to find the best possible correction [5] which is described in this chapter as well.

This chapter continues on reviewing parts of NLP tool tasks and how they are affected by the new environment of social media text. The tasks reviewed are Tokenizers, Part-of-Speech (POS) taggers, Chunkers & Parsers, and Named Entity Recognizers (NER). They are all measured using the stable of Precision, Recall, F-Score, and Accuracy but they all calculate those measurements in their own way. For instance, Chunkers & Parsers use a phrase-structure bracketing system defined by the Parseval Evaluation Campaign [6] and POS taggers are evaluated by looking at the number of correctly assigned words. They all suffered in performance when dealing with social media data. However, NER seemed to show the biggest decrease in performance losing almost 50% of performance when trying to recognize entities from formal edited text to the unruliness of social media text [7, 8, 9]. Thus some adaptations for NER include "microtext normalization" and more specialize NER algorithms for specific online platforms like Twitter. Other adaptations for the other task have their own variations but the general strategy is to specialize the task for a specific online platform, like adding twitter specific tags to training data sets. A very interesting adaption was regards to identifying the language of a given text in social media introduced by Carter [10]. Instead of simply analyzing data, they took advantage of the new environment of social media that included geo-location tags, user profiles, and threads of conversations in order to deduce the language. This was a great example of adapting well to the new data environment.

As what was mentioned before, data collection and annotation is the heart of the issue when dealing with this new environment of social media text. Chapter five of *Natural Language Processing for Social Media* [1], throughly discusses this issue. We have the developed ideas of NLP algorithms and if we had the right data we could easily adapt them to the new environments. Thus adapting the NLP methods can be quite tricky to get around the obstacle of having un-annotated data. We need good annotated data in order to train NLP models and definitely or testing purposes.

Social media has created a massive pool of data to analyze and collect. There are a few important features that we need to take into consideration when dealing with this data. There is public information which is readily available but some companies have limitations as to how much data you can actually acquire using their APIs without some type of payment subscription. Personally I developed a twitter scrapper for the Monterey Bay Aquarium and faced this issue when trying to extract millions of tweets. Secondly, there is private data associated with private user accounts which is un-collectable (at least ethically un-collectable). Lastly, validity of the text needs to be examined depending on the application but this is a big deal in the realm of social media. Since anyone can author text to the public associated to other people of entities like businesses, the validity of information can cause a significant effect. For instance this chapter references a statistics from a Harvard article [11] in which they found that about 20 to 30% of Yelp business reviews were fake. Additionally, the social media texts is disorganize and varies tremendously by incorporating slang and regional dialects from different countries down to neighboring cities.

Overall the realm of social media is exciting for Natural Language Processing and the challenges of collecting data has produced interesting problems as well as interesting solutions by researchers in this field.

References

[1] Farzindar AA, Inkpen D. Natural Language Processing for Social Media, Third Edition. Synthesis Lectures on Human Language Technologies. 2020;13(2):1-219. doi:10.2200/S00999ED3V01Y202003HLT046

[2] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. Nature, 521:436–44, 05 2015. DOI: 10.1038/nature14539 7, 16

[3] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 05 2017. 17, 78

[4] Seniz Demir. Context tailoring for text normalization. In Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing, pages 6–14, San Diego, CA, USA, June 2016. Association for Computational Linguistics. http://www.aclweb.org/anthology/W16-1402. DOI: 10.18653/v1/w16-1402 18

[5] Md Shad Akhtar, Utpal Kumar Sikdar, and Asif Ekbal. IITP: Hybrid approach for text normal- ization in Twitter. In Proceedings of the Workshop on Noisy User-generated Text, pages 106–110, Beijing, China, July 2015. Association for Computational Linguistics.

[6] Phillip G. Harrison, S. Abney, E. Black, D. Flickinger, C. Gdaniec, R. Grishman, D. Hindle, R. Ingria, M. Marcus, B. Santorini, and T. Strzalkowski. Evaluating syntax performance of parsers/grammars of English.

[7] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu- Sung Lee. Twiner: Named entity recognition in targeted Twitter stream.

[8] Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. Joint inference of named entity recognition and normalization for tweets.

[9] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study.

[10] Simon Carter, Wouter Weerkamp, and Manos Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text.

[11] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. Technical report, Harvard Business School