Natural Language Processing for Sentiment Analysis

Gabriel Hurtado, Vanessa Servais,
Melan Vijayaratnam, Paul Claret
{ghurtado6},{vservais3}, {mvijayaratnam},{pclaret3}@gatech.edu

This document presents a description of our project in the context of the Statistical Machine Learning course taught by Professor Anderson at Georgia Institute of Technology.

1 Summary

With the rise of machine learning as well as of the social media, Sentiment Analysis has grown rapidly. It focuses on computationally analyzing peoples sentiments, opinions and emotions in regards of products or other topics. There is a large spectrum to sentiment analysis and many ways in machine learning to approach it and its different applications.

For our project, we will focus on studying customer reviews, such as for products on Amazon, or commentaries on different topics like those on reddit. This analysis would be to translate or convert the written data into a quantified star like review.

Understanding customer satisfaction is crucial for businesses nowadays. Furthermore, it is the basis for political campaigns. Whenever you want to reach people, you have to understand what they think and how they react. However, it is very hard to work with all the tremendous amounts of written data that you get with customer reviews and comments. Therefore, it is very interesting and useful to quantify such data to analyze it fast and easily. If you read a review, you cannot easily say how much a customer liked a product, whereas a number on a scale tells you at the first look to what extent the customer is satisfied - or not. This quantification will speed up and simplify the job of data analysts in marketing: Working with quantitative data is simpler than qualitative one.

In our approach, we start with basic classification algorithms seen in class. We will then move to more complex ones using recurrent neural networks and deep learning. By using different algorithms, we are able to see how well they do on the datasets and which ones are more adapted to this problem. Furthermore, as the more complex algorithms do not only take into account individual words but the meaning of whole phrases, we will see the impact of these differences.

2 Description

2.1 Introduction

Sentiment analysis is the computational study of peoples opinions, sentiments, emotions and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics and many more. The inception and rapid growth of the field coincide with those of the social media on the Web, for example, reviews, forum discussions, blogs, Twitter and other social networks.

Accurate sentiment classification has many applications. It can help companies understand how customers feel about their products, politicians understand how people are responding to a policy proposal, or be used to investigate cultural differences (Pappas et al., 2016).

As such, effectively making use of text data in sentiment classification is valuable since it has a profound impact to business and society as a whole. This proliferation is due to the fact that opinions are central to almost all human activities and are key influencer of our behaviors.

2.2 State of the of art

Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP). Whilst this is a hard problem, significant progress has been made (Socher et al., 2013) using recurrent neural networks.

Essentially recurrent neural networks or RNNs (Rumelhart et al., 1986), are a family of neural networks for processing sequential data. This allowed to understand compositionality in tasks such as sentiment detection but as a consequence requires more evaluation resources.

The main issue with RNN comes from the fact it is not efficient at capturing long range connections which comes from vanishing gradient at the core. To remediate this problem, a modification to the RNN hidden layer was done and resulted in a new architecture called Gated Recurrent Unit (GRU) (Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling). Therefore it allows the neural network to learn even very long range dependencies.

On the same approach, an older architecture called Long Short Term Memory (LSTM), allows to learn very long range connections in a sequence (Hochreiter & Schmidhuber, 1997. Long short-term memory). It has been the historically more proven choice over GRU but trends are suggesting that in the last few years, GRUs have been gaining a lot of momentum.

Those architectures only took into account previous words for each word when performing computations. This is a fundamentally flawed reasoning since future words also have an influence to an actual word. This is what Bidirectional RNN (BRNN) are for; they allow taking information from both earlier and later of a sequence. Each hidden layer of the BRNN can then be used using GRU or LSTM although for a lot of NLP problems, a bidirectional RNN with LSTM appears to be commonly used.

2.3 Our approach and relevant points

Sentiment Analysis is a very broad field, many different angles could be chosen to apply interesting machine learning techniques in this area. Among all the possibilities, we will focus on the task of understanding the degree of satisfaction of a text, to translate it to a star review, as seen on many websites, such as Amazon or Imdb. This is a very trendy topic, as understanding customer feedback is pivotal in any commercial environment. Having the capability of understanding to what extend the customer is satisfied can provide a great feedback and gives room to the company for improvement. Even more, when this feedback can be quantified, so that it becomes very easy and not too time consuming to get a fast overall idea of the customers satisfaction. This technique can of course be applied to online reviews, but we can easily use such a training pattern with audio files converted to text, virtually widening even more the possible fields of application for such a problem.

This kind of problems can seem less interesting than the ones about less down to earth topics such as style transfer or video game reinforcement learning, but it has the advantage of being really needed and useful, as it provides better understanding of existing data, which is available in tremendous quantities, but of which they are not enough ways of extracting the knowledge.

Indeed, such an approach would make it much easier for companies or politicians to work with this type of written data. As our project is about quantifying sentiments and other rather difficult writings to easily categorize, it would allow them to work effectively with this data at their disposal. They could get the average satisfaction or opinion or even subdivide the data (and authors) into sub datasets to better target their future efforts (marketing, election campaigns,). Doing this type of sentiment analysis on comments simplifies the analysis of these vast amounts of data. Among other uses, we want to be available to convert a movie review commentary to a rating, as well as a product review, and we are considering to apply this same technique to some less expected data such as reddit commentaries or even tweets to companies.

The interest of this approach is that it is broad enough to include some interesting cases (such as the use of irony), using techniques that we did not cover in class, mainly focused on a Deep Learning approach, in order to have some kind of representation for the meaning and general intention behind the text. A key point on our interest on this problem is that it is at the same time doable even with simple techniques, such as the ones we discussed during class. Among those, we will try the Naive Bayes and Support Vector Machines (SVM) as well as the forthcoming ensemble method, that is random forests. Simple models cited previously will give us some kind of idea on how well our Deep approaches are doing, and giving us a sense on what is the gain of this increased hypothesis set in terms of error rate over test data. And thus, to see the impact of not only analysing each word solely but whole sentences or paragraphs.

This topic can be seen as a regression or classification problem, depending on if we choose to predict a real number, or just a class out of the 5 stars range. We agreed to see it as a classification problem for the time being, as it will make a Deep Learning approach easier to build, as we will only need output layers of length five, one for each possible rating. To be able to do so, we will use datasets that have already a quantification of the review or commentary to train our classifiers.

2.4 Workflow

We will start the project with some logistics. We will start the project with some logistics. Indeed, seeing that we have a small amount of time to complete the project, we will need to be efficient. In order to stay organized for the duration of the project, we will use different tools. For example, we will use a version control system and a project management application such as git and trello to keep everyone and all the files up to date. This means that we will need to learn how to use them, as quickly as possible. Said tools will ensure that everyone has all the information to get their part of the work done at all times, as well as to see the progress made in each of the tasks. This will allow the group to subdivide into smaller groups, and work on different tasks at the same time. Each sub-group will have a precise task and a leader to ensure the tasks completion on schedule.

In addition to those tools, we are also going to have at least a weekly meeting, in order to give feedback to everyone and to ensure that we are on schedule, and adapt the direction of the project if need be. Said meetings will be held every Thursday afternoon and will be mandatory

3 Collaboration Plan

Task	Reading the papers that are mentioned in the project proposal
Leader	Melan
Deadline	03/23
Importance	*** We need to do some research to understand the problem correctly.
Potential challenges	If the research is not done properly, the risk is high that the project will go in the wrong direction or that we wont understand how to implement the algorithms.

Task	Finding data for the training
Leader	Gabriel
Deadline	03/30
Importance	**** We need datasets to train our classifiers on and to later test them with test datasets.
Potential challenges	Without datasets, the implementation of the algorithms is impossible. Look for more datasets, even if they have not yet been cleaned.

Task	Data Cleaning
Leader	Melan
Deadline	04/19
Importance	****
Potential challenges	How to clean the data is an important issue. On the one hand, if we delete some sentences this will result in less performing algorithms. On the other hand, completing some unfinished sentences would mean that we introduce bias in our data so we have to be careful on that point.

Task	Feature engineering
Leader	Paul
Deadline	04/19
Importance	** Clearly identifying the important features in order to train correctly the classifiers.
Potential challenges	Learning to remove useless words such as prepositions from the data set. This involves the choice of the right stemmer from the Natural Language ToolKit (NLTK) library and other practices that come from linguists.

Task	Use of collaborative tools such as Trello and Git
Leader	Gabriel and Melan
Deadline	03/23
Importance	** Git and Trello tools are useful for group projects and to make the collaboration easier.
Potential challenges	Potential merge conflict when submitting code on Github as well as the effective use of branches to see contributions.

Task	Developing a classifier using Naive Bayes
Leader	Paul
Deadline	04/05
Importance	** This algorithm is important to get a first idea of how the classification will work.
Potential challenges	If the Naive Bayes approach does not work, we could only use the SVM technique, which could give us a first idea as well.

Task	Developing a classifier using either a SVM or an SVR
Leader	Vanessa
Deadline	04/05
Importance	** This algorithm will give us another angle to the problem or just fortify the Naive Bayes conclusion
Potential challenges	If the SVM (SVR) algorithm cannot be implemented, we would still have the Naive Bayes approach as a simple first classifier.

Task	Developing a recurrent neural network
Leader	Melan and Gabriel
Deadline	04/12
Importance	***
Potential challenges	To be able to generalize the system to new set of words, we might want to use a huge dictionary as well as word embeddings to get a representation or the words and this is computationally expensive

Task	Deep Recurrent Neural Networks using variants
Leader	Melan and Gabriel
Deadline	04/19
Importance	****
Potential challenges	A deep RNN can be built on top of a simple RNN and the variants such as BRNN with LSTM are taking multiple factors into account such as memory. Moreover, in common deep learning frameworks it is challenging for making them work without running into some errors that do not really help.

Task	Ensemble Methods like Random Forest, Gradient Boosting
Leader	Vanessa
Deadline	04/19
Importance	**
Potential challenges	Even if libraries such as sklearn, Xgboost, or even LightGBM make easy to experiment with such algorithms, it is easy to overfit the data when not knowing how to use the hyperparameters.

Task	Writing the final report
Leader	Paul
Deadline	04/24
Importance	High: The report is not actually necessary to realize the project but crucial for the grade.
Potential challenges	The algorithms take too much time to be implemented and run, so that we dont have the results in time or not the time to write the report. In this case, we could implement only the first algorithms and focus the report on those algorithms.