Milestone 2

Kevin De Angeli - Hector D. Ortiz-Melendez

November 12, 2019

BACKGROUND STUDY

Every day, millions of people post online their opinions about products, movies, places they visited, among many other activities and experiences. Together, this large amount of posts can be made into a dataset and fed to different machine learning (ML) and deep learning (DL) algorithms to understand people's general sentiment. In the context of e-commerce, customers' reviews can be useful for manufacturers because they tell what customer liked or disliked about a product [1]. However, you can now find hundreds or thousands of reviews for a single product, and inferring the general opinion of a large pool of comments is expensive and time consuming. Sentiment analysis is one of the tasks of natural language processing (NLP), a subfield of linguistics and computer science.

Scientist have approached sentiment analysis at different levels. Turney [2] developed an unsupervised learning algorithm to classify reviews as positive or negative. His work focuses on text analysis as a whole. Other authors have performed sentiment analysis at the sentence level. That includes Hu and Liu [1] who presented an algorithm to mine and summarize customer reviews of a product. However, unlike traditional text classification as a whole, their work focuses on identifying the features of the product on which the opinions are positive or negative. Recently, scientist have also been successful at performing sentiment analysis at the phrase level. For example Wilson et al. [3] presented a new approach for analysis at the phrase level to identify between neutral or polar expressions. Nevertheless, there are many more linguistic challenges to tackle such as ambiguous comments, specifically sarcasm identification where the sentiment is implied. Sarcasm is an interesting challenge given its topi-dependency and highly contextual nature. Expressions like sarcasm can be approached with techniques such as user profiling [4].

Another important challenges of NLP is feature space. Most of the earlier work in NLP was developed using an unigram (1-gram), bag-of-words (BoW) approach. BoW assumes that text is just a set of words where order does not matter; the corpus of text is then represented as a “document-term matrix of counts" [5]. This basic assumption about text has proven to be successful in NLP, and it has been applied to topic modeling and other tasks such as reporting partisan terms from political speeches [5]. Another popular feature extraction technique is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF basically tries to quantify how important a certain words is by observing how many times it appears in a specific document in comparison with how many times the word appears in the corpus [6]; words that appear frequently in a small number of documents but do not occur often in the whole corpus would have a high TF-IDF value. However, the apparent problem with unigrams is that they do not conserve the type of information that is inferred from the conjunction of two or more words. For example, phrases such as "social security" would completely lose their meaning under the BoW paradigm. One potential solution to this problem is $N-grams$. An N-gram refers to a sequence of ordered words where the length of the sequence is N. An $N-gram$ corpus provides information about how frequent a series of words occur [7]. Note that capturing the $N-grams$ of a corpus with dictionary size M will result in $M^N-grams$ [8]; in practice this number is prohibitively expensive, and we usually filter $N-grams$ which frequency are below certain threshold [7]. Another alternatives to the BoW model is parts-of-speech (POS). POS refers to mapping each word in a corpus to a specific part of speech (noun, pronoun, verbs, etc). This mapping process is known as part-of-speech tagging and is usually done with unsupervised learning algorithms that look into the contexts in which words occur to assign a specific label. POS are useful because they can provide a lot of information about a word and its neighbors [9]. The state of the art feature extraction technique is known as word embeddings. The idea behind word embedding is to "represent words as dense vectors that are derived by various training methods inspired from neural-network language modeling" [10]. One of the advantages of word embedding is that they attempt to capture word similarities. However, one significant drawback is that words can have multiple meanings, and that does not allow for some words to have a well-defined, single representation. One of the most popular word embedding algorithms is called Word2vec, and it was developed by a team lead by Tomas Mikolov at Google. Recently, Facebook has released a faster version of Word2vec called FastText. Since then, FastText has been used for numerous sentiment analysis tasks [11].

Scientists have developed sentiment analysis model for all kind of applications. Nogueira dos Santos et al. [12] have analyzed Twitter posts. This a challenging task because messages don’t provide much information about the context. They used a deep convolutional network (traditionally implemented for image processing tasks) and obtaiened an accuracy of 85.7%. Wöllmer et al. [13] analyzed the general sentiment of online videos by considering not only textual information but also audio features. Thet et al. [14] proposed a method of automatic sentiment analysis of movies reviews that provides orientations (positive or negative) and different strength of these orientations. Gräbnera et al. [15] implemented a lexicon-based approach to classify hotel reviews. They obtained their data from the TripAdvisor website. Finally, aggression identification is an extreme application of sen-

timent analysis that has been enabled by cyber-harassment and cyber-bullying in social media [16].

The dataset used for our paper contains sentences labeled with positive and negative sentiment. The data was originally collected by Kotzias et al. [17]. In their work, they used the dataset to test their new algorithm (GICF). Their results are presented in Table 0.1. They have only reported the accuracy of three algorithms, and two of these are logistic regression with two different feature spaces. In this project we will train multiple ML algorithms: Gaussian classifiers, kNN, multiplayer perceptron, decision trees, and support vector machine (SPM), and compare the accuracy results.

Model	Amazon	IMDb	Yelp
Logistic regression with BoW	79.086.3%	76.286.3%	75.186.3%
Logistic regression with Word Embeddings	54.386.3%	57.986.3%	66.586.3%
GICF with Word Embeddings	88.286.3%	86.086.3%	86.3%

Table 0.1: Kotzias et al. results for the Amazon, Yelp, and IMDB reviews dataset. They called they novel model "GICF".

REFERENCES

[1] Minqing Hu and Bing Liu *Mining and Summarizing Customer Reviews* pdf.

[2] Peter D. Turney *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. pdf.

[3] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*. pdf.

[4] Elvis *https://medium.com/dair-ai/detecting-sarcasm-with-deep-convolutional-neural-networks-4a0657f79e80*

[5] Abram Handler, Matthew J. Denny, Hanna Wallach, and Brendan O'Connor *Bag of What? Simple Noun Phrase Extraction for Text Analysis*. pdf.

[6] Juan Ramos, *Using TF-IDF to Determine Word Relevance in Document Queries*. pdf.

[7] Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale *New Tools for Web-Scale N-grams*. pdf.

[8] P Majumder, M Mitra, B.B. Chaudhuri *N-gram: a language independent approach to IR and NLP*. pdf.

[9] Daniel Jurafsky, James H. Martin *Speech and Language Processing*. Third edition.

[10] Omer Levy and Yoav Goldberg *Dependency-BasedWord Embeddings*. pdf.

[11] I. Santos, N. Nedjah and L. de Macedo Mourelle *Sentiment analysis using convolutional neural network with fastText embeddings*, 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Arequipa, 2017, pp. 1-5. doi: 10.1109/LA-CCI.2017.8285683

[12] Ciccero Nogueira dos Santos, and Maria Gatti *Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts*. pdf.

[13] Martin Wöllmer, Felix Weninger, Tobias Knaup, and Björn Schuller, Technisch *YouTube Movie Reviews: Sentiment Analysis in an AudioVisual Context*. pdf.

[14] Tun Thura Thet, Jin-Cheon Na and Christopher S.G. Khoo *Aspect-based sentiment analysis of movie reviews on discussion boards*. pdf.

[15] Dietmar Gräbnera, Markus Zankerb, Günther Fliedlb, and Matthias Fuchs *Classification of Customer Reviews based on Sentiment Analysis*. pdf.

[16] Raiyani, Kashyap, Gonçalves, Teresa, Quaresma, Paulo, and Nogueira, Vítor *Fully Connected Neural Network with Advance Preprocessor to Identify Aggression over Facebook and Twitter*

[17] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth1 *From Group to Individual Labels using Deep Features*. pdf.