VERSION 1.3

OCTOBER 26, 2017

STORY NEXT 2.0

A TEXT INSIGHTS/VISUALIZATION TOOL

WHAT **STORY** DOES YOUR CONTENT SAY?

BISWARAJ KAR

ANDREW KRISCHER

LING ZHANG

NORTHEASTERN UNIVERSITY

COLLEGE OF COMPUTER AND INFORMATION SCIENCES

PROJECT PROGRESS REPORT FOR **NATURAL LANGUAGE PROCESSING (CS 6120)**

STORY NEXT 2.0

INTRODUCTION

This project aims to help *content-writers* uncover subtle patterns in their creative work, such as potential biases and overall sentiment. The overall idea is to create a **visual analytical tool** that processes written content to give unique insights. Our project goal is to have a visualization which will show sentiment of a text in an elegant way so that it gives a clear picture of the sentiment of the written text/article/story at a glance and enable the user to get interesting insights just by looking at the visualization. The intended audience for the project are story tellers, script writers, journalists, creative-writers and content consumers like us.

CHANGES AND TASK MODIFICATIONS

Building upon what we wrote in the proposal, we narrowed down the focus of the current work to only consider and build a model around a binary class sentiment of [Positive, Negative] as opposed to considering *Neutral* as a 3rd class for sentiments. The reason for doing is the 2 fold:

- The domain we are targeting (News & Literature) is very new when it comes to application of sentiment analysis or opinion-mining and is an area of active research. As a result, we were not able to find any annotated datasets for our domains to train our model in. That would mean that we have to create own big corpus for training, by manually tagging documents/texts. Given the scope and timeline, we think a 3 sentiment classification is not a feasible idea for now and we will explore this further once we have a stable data-model later in the project.
- We found that the contextual information to decide *Neutrality* of opinion is not very straight forward in large texts like news articles and literature, which have very high objectivity often. Interestingly, previous attempts like classifying movie reviews ([1]*Pang and Lee 2002*) is relatively a less complex task as the "Star Rating" indicators can easily give scale to ascertain neutrality (e.g. on a 0-9 rating scale, [1-3][4-6][7-9] can easily be assumed to be [Neg][Neu][Pos]). Similarly, for tweets or blogs, where the absence of Emoticons can imply neutrality directly ([2]Yang et al.); which in our case, doesn't apply and is a much more complex task due to size and scope of news-articles/literature.

DATA PROCESSING

Due to the non-availability of an annotated training data after an extensive search of more than 10 publications and many research corpuses (Stanford, Brown, UPenn, UCI ML Repository, Kaggle etc.) we decided that we will attempt classification of our domains from language-models trained from the Movie Corpus of ([1]*Pang and Lee 2002*). Our research revealed that the only reliable pre-annotated large sentiment corpora are only available in specific domains, like Twitter, Movie Reviews, Product Reviews on Amazon and Blogs based on emoticons. We chose the Movie Corpus among all the options to train our model due to some inherent intuition of the movie review corpus:

- The movie review corpus has diverse data that is comprised of one or many paragraphs of user opinion; classification is not bound by size or specificity of features as in the case of Twitter (140 characters) or presence of emoticons in blogs.
- The movie review corpus is sufficiently large (9,645 sentences) to achieve some generality of natural language, given they are actual human opinions. The size and depth of the corpus also gives it a little bit more domain independence.

- Many other successful sentiment analysis mechanisms, like the Stanford NLP ([5]Socher, Richard, et al.) also use the Movie review ([1]*Pang and Lee 2002*) corpus and yet are more widely applicable to any domain using deep learning. Though they employ more advanced techniques like Recursive Neural Networks and Deep Learning, the inherent base learning model is indeed the same movie review corpus, which calls for some merit for using the corpus.

Hence, for our sentiment classification, we decided to begin with the movie review dataset ([1]*Pang and Lee 2002*) that was given to us in assignment two. The corpus contains already-tagged sentiment data, making it a perfect candidate for training in our early stages of the project.

METHODOLOGY

Extending our intuition and plan in the proposal, we've used the Unigram and Bigram models for feature representation of our data. We calculated and stored the unigram and bigram counts from the movie review corpus and then calculated the probabilities for each word in the web page interface. The core feature for our project is classification. As a first step for classification, we used Naive Bayes classifier to train on the movie review corpus. We did not transform the training data too much as the training data is already classified and annotated from the dataset ([1]*Pang and Lee 2002*). Since we intend to use a different dataset to see if our Naive Bayes is working on different data types, we used some hand-tagged literature data (detailed in 'Evaluation' section) and some hand-tagged News articles from various sections. To clean our data, we first imported the data using the NLTK python library and then removed any empty sentences. We then tabulated unigram and bigram counts and stored them as dictionary in PKL format. This allows us to easily reload and reuse the preprocessed data for any algorithm we choose in the future. We then ran our models on the hand classified Literature and News articles and evaluated the results.

EVALUATION

We did a mix of Intrinsic and Extrinsic evaluations to measure the success of the project. For the extrinsic part, we tested our classifier on a set of real stories and news articles which will be hand-annotated by us.
Additionally, we employed some standard techniques like computing accuracy ([3]Manning and Schutze, 1999) of the classifier on the whole evaluation dataset, defined as:

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad , \quad recall = \frac{True\ Positives}{True\ Positives + False\ Negative}$$

Where a $True\ Positive$ implies a text being actually of positive sentiment and the classification also gives a positive sentiment label to the text. $False\ Positive$ implies that implies a text being actually of negative sentiment and the classification also gives a positive sentiment label to the text. Similarly, the $False\ Negative$ measure was also derived. To measure the performance, we also used F-measure ([3]Manning and Schutze, 1999), which is the harmonic mean of precision and recall:

$$F = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2\ precision + recall}$$

Intrinsically, we took 70% of our dataset as testing data and we got more than 83% accuracy on movie reviews which is intuitive given the training was also trained on the Movie corpus. For our cross-domain testing, the available corpora with sentiment tags was limited to few domains (as we discussed in the 'Data Processing' section), we were interested in how effective our classifier works with cross-domain data. We employed on testing as the following:

LITERATURE TESTING:

Using the **Gutenberg corpus**, we selected 24 sentences from Jane Austen's *Emma* *[L1 in Corpus References]*. We then hand-tagged each of the 24 sentences with a positive or negative sentiment. In total, by chance, there were 12 positive and 12 negative hand-tagged sentences. We ran the Naive Bayes classifier (that was trained on movie review data) on those sentences, and came up with the following results:

	Predicted Positive	**Predicted Negative**
Actual Positive	11 (TP)	1 (FN)
Actual Negative	6 (FP)	6 (TN)

This results in a ***Precision*** of 65%, **Recall** of 91.7% and a **F1** of 75%. This indicates that there may be some innate efficacy between literature and movie review domains.

NEWS TESTING:

Using the articles from the New York Times, Guardian, CNN and the Huffington post *[N1-12 in Corpus References]*, we selected 12 news articles from Politics, Entertainment, Travel and Sports. We then hand-tagged each of the 12 sentences with a positive or negative sentiment. In total, by chance, there were 7 positive and 5 negative hand-tagged sentences. We ran the Naive Bayes classifier (that was trained on movie review data) on those sentences, and came up with the following results:

	Predicted Positive	**Predicted Negative**
Actual Positive	4 (TP)	3 (FN)
Actual Negative	3 (FP)	2 (TN)

This results in a ***Precision*** of **57%**, **Recall** of **57%** and a **F1** of **57%**. This indicates that for news, corss training might perhaps not be a great idea and we need to explore better methods if we are to improve the scores.

NEXT STEPS

As next steps to proceed, we are working on the following approaches:

1. Implementing classification using SVM instead of Naïve Bayes as SVM does not assume class independence and has been proven to be more accurate in previous works[1,2,3,4].
2. Using MPQA's (http://www.aclweb.org/anthology/N/N15/N15-1146.pdf) opinion corpus's affect based methods[9][10] instead of B-/Tri-gram language models.
3. Several researchers have suggested an alternative method to the use of dictionaries that report the sentiment of a set of words along one or more emotional dimensions. Examples of sentiment dictionaries includes POMS[8] and POMS-ex—Profile of Mood States—and ANEW—Affective Norms for English Words. We are working on using ANEW, an extended ANEW dictionary[6] that was recently built by researchers McMaster and Ghent Universities, and a happiness dictionary[7] built by researchers at the University of Vermont. These are basically linguistic models based on "Valence and Arousal"[6] to tag sentiments. We will report the accuracy of this approach in our next status report.

REFERENCES

Technique/Methodology References:

[1] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02), Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 79-86. DOI: https://doi.org/10.3115/1118693.1118704

[2] C. Yang, K. H. Y. Lin and H. H. Chen, "Emotion Classification Using Web Blog Corpora," Web Intelligence, IEEE/WIC/ACM International Conference on, Fremont, CA, 2007, pp. 275-278.

[3] Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

[4] Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 417-424. DOI: https://doi.org/10.3115/1073083.1073153

[5] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

[6] Norms of valence, arousal, and dominance for 13,915 English lemmas. Warriner AB, Kuperman V, Brysbaert M. Behav Res Methods. 2013 Dec;45(4):1191-207. DOI: 10.3758/s13428-012-0314-x. PMID: 23404613

[7] Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE6(12): e26752. https://doi.org/10.1371/journal.pone.0026752

[8] L. Curran, Shelly & A. Andrykowski, Michael & Studts, Jamie. (1995). Short form of the Profile of Mood States (POMS-SF): Psychometric information. Psychological Assessment. 7. 80-83. 10.1037/1040-3590.7.1.80.

[9] Yoonjung Choi and Janyce Wiebe (2014) +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference, Proc. of EMNLP 2014.

[10] Janyce Wiebe, Theresa Wilson , and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.*

Corpus References:

[L1] Jane Austen's Emma http://www.gutenberg.org/files/158/158-h/158-h.htm

[N1] Boko Haram strapped suicide bombs to them. Somehow these teenage girls survived https://www.nytimes.com/interactive/2017/10/25/world/africa/nigeria-boko-haram-suicide-bomb.html?hp&action=click&pgtype=Homepage&clickSource=story-heading&module=second-column-region&region=top-news&WT.nav=top-news

[N2] Trump Falsely Denounces Jeff Flake by Calling Him a ... Democrat? https://www.nytimes.com/2017/10/25/us/politics/fact-check-trump-flake-democrat.html?_r=0

[N3] Albert Einstein's 'Theory of Happiness' Fetches $1.56 Million https://www.nytimes.com/2017/10/25/world/middleeast/einstein-theory-of-happiness.html?rref=collection%2Fsectioncollection%2Fscience&contentPlacement=2&module=Slide&region=Sl

ideShowTopBar&version=SlideCard-1&action=Click&contentCollection=Fashion%20%26%20Style&slideshowTitle=Mariah%20Carey%20Hits%20Her%20High%20Notes%20at%20Karl%20Lagerfeld%20DInner&currentSlide=1&entrySlide=1
[N4] 'Tyler Perry's Boo 2' Is No. 1 Amid a Plethora of Duds
https://www.nytimes.com/2017/10/22/movies/tyler-perrys-boo-2-madea-halloween-box-office.html?rref=collection%2Fsectioncollection%2Fmovies&action=click&contentCollection=movies&region=rank&module=package&version=highlights&contentPlacement=2&pgtype=sectionfront
[N5] With a Crowd of Diverse Faces, Dodger Stadium Stands Out
https://www.nytimes.com/2017/10/25/sports/baseball/los-angeles-dodgers-world-series.html?rref=collection%2Fsectioncollection%2Fbaseball&action=click&contentCollection=baseball&region=stream&module=stream_unit&version=latest&contentPlacement=3&pgtype=sectionfront
[N6] World Series 2017: How the Astros Won Game 2, Inning by Inning
https://www.nytimes.com/2017/10/25/sports/world-series-dodgers-astros.html?ribbon-ad-idx=3&src=trending&module=Ribbon&version=origin&region=Header&action=click&contentCollection=Trending&pgtype=article
[N7] George H.W. Bush Apologizes After Actress Says He Sexually Assaulted Her
https://www.huffingtonpost.com/entry/george-h-w-bush-actress-sexual-assault_us_59f05410e4b0bf1f8836dea0?ncid=inblnkushpmg00000009
[N8] Worried Trump Will Strike North Korea, Democrats Pitch Bill To Slow Him Down
https://www.huffingtonpost.com/entry/democrats-trump-north-korea-bill_us_59f14ba7e4b07d838d320ed3
[N9] Escape to China's Land of the Yellow Dragon
https://www.greatbigstory.com/stories/escape-to-china-s-land-of-the-yellow-dragon?playall=1417
[N10] Paul Ryan vows that help is coming after hurricane hit Texas
http://www.cnn.com/2017/09/21/politics/paul-ryan-houston-visit-hurricane-aid/index.html
[N11] NAACP Warns African-Americans Against Traveling On American Airlines
https://www.huffingtonpost.com/entry/naacp-travel-advisory-american-airlines_us_59f0bf2fe4b03c73bf347869
[N12] All five living former US presidents make rare appearance together
https://www.theguardian.com/us-news/2017/oct/22/all-five-living-former-us-presidents-make-rare-appearance-together.