VERSION 1.2

OCTOBER 2, 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 PROPOSAL FOR **NATURAL LANGUAGE PROCESSING (CS 6120)**

Biswaraj Kar | Andrew Krischer | Ling Zhang
github.com/biswarajkar/storynext2

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. Additionally, we believe we can use the work done in this project to aid the audience in consuming the content too. The overall idea is to create a **visual analytical tool** that processes written content to give unique insights. This is potentially a part of a larger future project for converting written content into a 3D generated visual story, viewable on a VR platform. For our current project, however, the intended audience are story tellers, script writers, journalists, creative-writers and content consumers like us.

MOTIVATION

The motivation for the project stems from interviews we did with writers and journalists, and their common desire for making story-writing and story-telling more introspective and insightful. Our interviews illuminated problems that NLP research may able to address. We condensed the suite of issues into a few questions, to aid the direction of our project:

- What if there was a tool that shows what sentiments your writing conveys, as you are composing it?
- What if we condense a 20-page story into one a one-page interactive visualization, conveying the major themes and motifs in a succinct manner?
- In today's globalized world, can we pre-empt writers to write more culturally sensitive material by automatically hinting culture-specific context (which they might not be aware of)?

In the quest to find answers to the above questions, we plan to work on this project to use visualizations and NLP to *create a tool that aids content-writers* ***during*** *the process of content-generation*. We believe this will not only empower *authors* to write more engaging, relevant and balanced content, but will also help the general *audience* gauge if a writing piece interests them.

RELATED WORK

There have been many papers written on sentiment analysis for the domain of blogs and product reviews. (Pang and Lee 2008) give a survey of sentiment analysis. Researchers have also analyzed the brand impact of microblogging (Jansen). We could not find any papers that analyzes NLP techniques in the specific domain of story writing or journalism, probably because digital story-telling and real-time opinion mining is a relatively new field. Overall, text classification using machine learning is a well-studied field (Manning and Schuetze 1999). (Pang and Lee 2002) researched the effects of various machine learning techniques (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM)) in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9% using SVM and a unigram model. Researchers have also worked on detecting sentiment in text. (Turney 2002) presents a simple algorithm, called semantic orientation, for detecting sentiment. (Pang and Lee 2004) present a hierarchical scheme in which text is first classified as containing sentiment, and then classified as positive or negative.

DATASETS

The primary idea is to derive sentiment insights from any text or story inputted by the user. To achieve a high quality sentiment prediction, we will need to train our language model on a large corpus of literature, news and historical texts and apply them to the test data as per the domain. Some of the primary training datasets we plan to use for the project are outlined below:

- Literature Corpora: we plan to use the following corpora to train our language models on books, novels, history and literature:
  - Gutenberg Corpus in NLTK

  - Google Books Ngram Dataset: storage.googleapis.com/books/ngrams/books/datasetsv2.html
  - Open American National Corpus: anc.org/data/oanc/download/
  - Brown Corpus: clu.uni.no/icame/manuals/BROWN/INDEX.HTM
- News/Journalism Corpora: we plan to use these corpora to train our language models on news articles, editorials and journalistic material:
  - BBC (2225 articles) and BBC Sports (737 articles) Corpora: mlg.ucd.ie/datasets/bbc.html
  - Reuters Full Dataset (2007-current): reuters.com/resources/archive/us
  - Library of congress Chronicling America (1798-1925): chroniclingamerica.loc.gov/about/api/
  - The NYT Annotated Corpus (1987-2017): catalog.ldc.upenn.edu/LDC2008T19

Our test data will be live data typed through a web-based interface, or a text file uploaded by the user.

METHODOLOGY

Given the goal and the data we have, we plan to evaluate a few methodologies and moving forward with the highest quality ones. Something that we plan to do differently is to classify documents by 3 sentiments- Positive, Neutral and Negative; rather than the binary Positive and Negative sentiments. Sentiment analysis using binary classifiers has been done extensively; we wish to see if we can expand on it. We believe collecting a substantial number of neutral stories will be very challenging but also necessary. We outline the key elements of the methodology as the following:

User Input/Data Collection

There are not any existing data sets of classified stories for analysis. Our test data will be live data typed by the user through a web-based interface or by uploading a text file which has the story/text.

Training Data

As outlined in the 'Datasets' section, our training data will be relevant corpora from different domains.

Classifiers

We plan to code a modified Naive Bayes classifier and use 3rd party libraries for trying Maximum Entropy and Support Vector Machines. Based on the outcome, we plan to refine or drop or change our choice of classifiers.

Feature Extractors

For now, we are hoping to try out the simplest Unigram and Bigrams as we wanted to smooth out instances like 'not good' or 'not bad' as negation is a vital element for sentiment analysis. We also plan to use Part of Speech (POS) features as they clarify homographs. For example, 'over' as a verb might have a negative connotation whereas 'over' as the noun, would refer to the cricket over which by itself doesn't carry any negative or positive connotation.

EVALUATION

We plan to do a mix of Intrinsic and Extrinsic evaluations to measure the success of the project. For the extrinsic part, we plan to test our classifier on a set of real stories and news articles which will be hand-annotated by us. Additionally, to evaluate intrinsically, we intend to employ some standard techniques like computing accuracy (Manning and Schutze, 1999) of the classifier on the whole evaluation dataset and compute accuracy across the classifier's decision (Adda et al., 1998), defined as:

$$decision = \frac{N(retrieved\ documents)}{N(all\ documents)}$$

Where N is the number of sentiments. The value of the decision shows what part of data was classified by the system. To measure the performance, we plan to use F-measure (Manning and Schutze, 1999), which is the harmonic mean of precision and recall:

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