Instructions for NAACL HLT 2016 Proceedings[[1]](#footnote-2)\*

|  |
| --- |
|  |
|  |
|  |
| Anonymous NAACL submission |
|  |
|  |
|  |
|  |
|  |

Abstract

This report presents our approach towards stance analysis of twitter data. Our task includes two sub tasks based on supervised learning and weak supervised learning. There are three labels Favor, Against and Neutral under which tweets are to be classified. To address these tasks, we have employed data pre-processing to filter the tweets, feature extraction and classification. The report presents preliminary results and will be improved by implementing future work.

Introduction

Social networking is becoming very prevalent in present scenario, where people express their opinions about different issues on platforms like Twitter, Facebook etc. These opinions can target anyone and can be in favor, against or may be neutral to that specific target. If the opinion or stance on a tweet completely supports the target entity the stance is considered as *Favor*, and if it completely berates the target entity the stance is *Against.* Also in some cases the stance is not explicitly reflected, these can be classified as *Neutral*. Great number of researchers are focusing on stance analysis to contribute in growth of natural language processing.

Stance analysis is often confused with sentiment analysis, despite having different approach. Sentiment represent an emotion of a person about a specific target i.e. it is basically how a person feels about a specific target, either positive or negative. Whereas stance is much broader concept which include sentiment as well as evaluation of information provided by that person. Stance always require a specific target entity about which the opinion is expressed, the target aspect of the entity and the sentiment associated with the stance. Detecting the correct stance from a tweet is a different challenging task.

The task we address is about classifying the given tweets under three labels Favor, Against or Neutral. Task is divided in to two sub tasks where one involves supervised learning other uses weak-supervised learning. For the first task we are provided with training data having five different target entities and the tweets have predefined labels for stance. The five target identities are *Hillary Clinton, Atheism, Climate change is a real concern, Feminist movement, Legalization of abortion.* For the second subtask we have only one target entity *Donald Trump*, weak-supervised learning will be implemented on a huge unlabeled data to detect the correct stance for the target entity. Our goal is to use the provided training data for both subtasks to train our model for stance detection.

Initially we will start our approach by preprocessing the data by removing all the disturbance from tweets in our training data. Then features selection will decide on the different features to be considered. After feature selection a classification algorithm will be applied to classify the tweets for its respective stance. At last we will evaluate the performance of our model by computing the f-score and compare the results with similar research.

Preprocessing Data

On social networking platforms like Twitter people usually do not follow the formal rules of a language and instead, use abbreviations, slangs which might generate deviations from the correct meaning of the tweet. Since computers are not capable of understanding these deviations so abbreviations and slangs must be corrected. Also people like to use *Hashtags*(#) between the or after the tweets and often tag different users by using *‘@’* followed by username. These hashtags and tags can again increase the complexity in determining the stance of the tweet. In our model we will eliminated the tags(@) and hashtags from the tweets, but we will collect the list of all hashtags associated with each tweet separately to use them as features for our classification algorithm.

|  |
| --- |
| Ruler |

Separating Hashtags

The hashtags are used along with tweets to highlight a specific topic of interest or to emphasize on a keyword currently in trend. Searching the hashtags on social media fetches all the information associated with those hashtags. Hashtags usually keywords preceded by ‘#’ character. Some of the hashtags in out training data are *#Benghazi, #SemST, #socialism, #WhyIamNotVotingForHillary* etc. Training model might encounter trouble in interpreting the correct meaning of them. We separated all the hashtags from tweets and collected them separately along with the unique Ids of each tweet.

Also we often see few words in a tweet preceded by ‘@’ character, these are called tags. These tags are used by people to notify other people that they are supposed to see that tweet. These also might degrade the prediction of classification. We have removed these tags from the tweets and eliminated them completely.

Eliminating Punctuations

The tweets in our training set also had many unnecessary and incorrect punctuations including multiples use of same symbol (“!!!!”, “…”). We removed all the punctuations from the tweet to clean the tweets. If there is an apostrophe(“ ‘ “) character, it will be identified during abbreviation removal step. For example, *it’s* after punctuation removal will become *its*. But after abbreviation and slang removal it will be transformed into *it is.*

Correcting Slangs & Abbreviations

Due to word character limit in Twitter, people are restricted to use a limited number of characters to express their opinion. Failing to do that, most of them have started using abbreviations of different words, so that they can write less but express more. For example, instead of *with respect to* we might see abbreviation *wrt*. Users also prefer to use slangs instead of using formal form of the word. For example *great*  is often written as *gr8*, *government* is written as *gov’t* etc. The stance detection method is not capable enough to recognize the correct form of these words.

The proposed model identifies these slangs and abbreviations in the tweets and translate them to its correct form. We have used a pre-defined dictionary1 of which consists of slangs/abbreviation and their respective correct form to perform this step.

# Classification

The classification process consists of feature selection, feature extraction and finally using machine learning algorithm to classify the given tweets in the testing data into the pre-specified classes. We have taken ideas from the existing literature and added a few concepts of our own.

Feature Selection

We have learned from the papers that we read that combining lexical and semantic features can help the system perform better. Sentiment features have also been used before. Our approach is to handpick a handful of topics that are closely related to our targets and use the associated sentiments of the overall tweet along with them. As lexical features we have tested using both bag of words and word bigrams. We are working towards extracting the semantic features as well.

Feature Extraction

We have taken the clean tweets from the preprocessing phase to extract features from them. For the first step of our classification process we are using word bigrams as the feature. For that purpose, we have first extracted all the bigrams from the cleaned training set tweets. Because there are many infrequent bigrams we have decided not to use them as features. Rather, we have taken the most common 500 bigrams from the training set data to build our feature set.

**3.3 Passing Data through Classifiers**

Although the stance detection task is about classifying the tweets into ‘Favor’ & ‘Against’ class, in the training data there are tweets that have the label ‘None’ also. Many of the literature we went through first classified between tweets with stance and those with none. Then they went ahead and classified the tweets with stance into Favor and Against. At this point in our project we have classified into the three classes. Later we will introduce a two-layer classifier.

We have used the Multinomial Naïve Bayes & the Support Vector Machine classifiers provided by the scikit-learn package wrapped by the nltk package.

# Preliminary Results

**We have performed the classification task on**

**Citations:** Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972).

**References:** References should appear under the heading **References** at the end of the document, but before any Appendices, unless the appendices contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a reference as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Authors’ full names rather than initials are preferred. You may use **standard** abbreviations for conferences[[2]](#footnote-3) and journals[[3]](#footnote-4).

**Appendices:** Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

**Acknowledgment** sections should go as a last (unnumbered) section immediately before the references.

Figures and Tables

**Creating:** To create new figures or tables, copy Figure 1, then replace the content and caption with the proper text. To make sure to capture all formatting, it may be wise to capture one line before and one line after the example as you copy it and then deleting the extraneous material. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

|  |
| --- |
|  |
| Figure 1: Figure caption. |

In MSWord, authors can place the graphic and its captions inside the rows of a 2×1 table (2 rows and 1 column) with invisible borders. Specify table positioning by right-clicking its handle in the upper left corner. Place the image in the center of the first row, and the caption in the center of the second row.

**Placing:** Place figures and tables in the paper near where they are first discussed, as close as possible to the top of their respective column. Wide illustrations may run across both columns and should be placed at the top of a page.

**Captions:** Provide a caption for every table and figure; number each one sequentially in the form: “**Figure 1:** Figure caption.”, “**Table 1:** Table caption.” Type the captions of the figures and tables below the body, using 9 point text. Table and Figure labels should be bold-faced.

|  |
| --- |
| Ruler |

**Numbering:** To update the numbering, press **Ctrl-A + F9**. This will update all the numbering applicable to tables, figures, equations, and headings.

**Cross-referencing:** To add a cross reference to a figure or table:

* Place the mouse pointer at the location where you wish to add the cross reference.
* Click on the **Insert** menu, (then click **Reference**), and then **Cross-reference** in the **Links** panel.
* In the **Cross-reference** dialog box, click the caption to which you are building the text reference.
* For a figure, under **Reference Type**, click **Figure**.
* Under Insert Reference To, click Only Label and Number, then click OK.
* Once the reference is in place, apply the ‘Normal’ font style (size 11, no bold face).
* This is a reference to Figure 1.

Equations

An example equation is shown below:

(1)

To add new equations, authors are encouraged to copy this existing equation line, and then replace with the new equation. The numbering and alignment of equation line elements is automatic. To update equation numbering, press **Ctrl-A + F9**. Note: this will only update the number to the right of the equation; to update numbering within the text you must create a cross reference.

**Cross-referencing:** To create a cross reference for an equation:

* Create a bookmark for it.
* Select the equation number to the right of the equation. Go to **Insert**, **Bookmark** (in the **Links** panel),andthen create a name for your equation. Press **Add** to create the bookmark.
* To refer back, place the mouse pointer at the location where you wish to add the cross reference.
* Go to **Insert, Cross-reference** (in the **Links** panel).In the dialogue box, select **Bookmark** and **Bookmark Text** from each dropdown list. Uncheck **Insert as Hyperlink**, then click **OK**.
* This will make it such that whenever a new equation is added, the references to the equation will be updated when **Ctrl-A + F9** is pressed.
* This an example cross-reference to Equation (1).

Footnotes

**Footnotes:** Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[[4]](#footnote-5) Footnotes should be separated from the text by a line.[[5]](#footnote-6) Footnotes should be in 9 point font.

|  |
| --- |
| Ruler |

From Submitted to Camera-Ready

To remove submission formatting for the camera-ready document, delete the header and footer on the first page (double clicking in the area, then delete the text). This will remove them from the rest of the document. To remove the ruler, click on Styles, the ACL Submission Ruler dropdown, “Select All N Instance(s)”. Then press Delete on your keyboard.

Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions.

Length of Submission

The NAACL HLT 2016 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content, plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page (up to 9 pages with unlimited pages for references) so that reviewers’ comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references. For both long and short papers, all illustrations and appendices must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Type of Text** | **Font Size** | **Style** | | paper title | 15 pt | **bold** | | author names | 12 pt | **bold** | | author affiliation | 12 pt |  | | the word “Abstract” | 12 pt | **bold** | | section titles | 12 pt | **bold** | | document text | 11 pt |  | | abstract text | 10 pt |  | | captions | 9 pt |  | | caption label | 9 pt | **bold** | | bibliography | 10 pt |  | | footnotes | 9 pt |  | |
| Table 1: Font guide. |

Double-blind review process

As the reviewing will be blind, the paper must not include the authors’ names and affiliations. Further- more, self-references that reveal the author’s identity, e.g., “We previously showed (Smith, 1991) ...” must be avoided. Instead, use citations such as “Smith previously showed (Smith, 1991) ...” Papers that do not conform to these requirements will be rejected without review. In addition, please do not post your submissions on the web until after the review process is complete (in special cases this is permitted: see the multiple submission policy below).

We will reject without review any papers that do  
not follow the official style guidelines, anonymity  
conditions and page limits.

Multiple Submission Policy

Papers that have been or will be submitted to other meetings or publications must indicate this at submission time. Authors of papers accepted for presentation at NAACL HLT 2016 must notify the program chairs by the camera-ready deadline as to whether the paper will be presented. All accepted papers must be presented at the conference to appear in the proceedings. We will not accept for publication or presentation papers that overlap significantly in content or results with papers that will be (or have been) published elsewhere.

Preprint servers such as arXiv.org and ACL-related workshops that do not have published proceedings in the ACL Anthology are not considered archival for purposes of submission. Authors must state in the online submission form the name of the workshop or preprint server and title of the non-archival version. The submitted version should be suitably anonymized and not contain references to the prior non-archival version. Reviewers will be told: “The author(s) have notified us that there exists a non-archival previous version of this paper with significantly overlapping text. We have approved submission under these circumstances, but to preserve the spirit of blind review, the current submission does not reference the non-archival version.” Reviewers are free to do what they like with this information.

|  |
| --- |
| Ruler |

Authors submitting more than one paper to NAACL HLT must ensure that submissions do not overlap significantly (> 25%) with each other in content or results. Authors should not submit short and long versions of papers with substantial overlap in their original contributions.

STREAM Tools

This Microsoft Word file has been preset for compatible use with the STREAM Tools template designed for creating well-formatted reports and papers with Microsoft Word. The principles behind this template and others STREAM templates are explained in (Mamishev, 2010; Mamishev, 2013).

Acknowledgments

Do not number the acknowledgment section.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J.Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114-133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

1. [↑](#footnote-ref-2)
2. https://en.wikipedia.org/wiki/  
   List\_of\_computer\_science\_conference\_acronyms [↑](#footnote-ref-3)
3. http://www.abbreviations.com/jas.php [↑](#footnote-ref-4)
4. This is how a footnote should appear. [↑](#footnote-ref-5)
5. Note the line separating the footnotes from the text. [↑](#footnote-ref-6)