**Do Established Companies Use Social Media Differently than Fast-Growing Companies?**

Aydan Alyanak

Indiana University   
School of Informatics and,

Computing, Data Science  
Bloomington, IN

alyanaka@indiana.edu

Derrick Eckardt

Indiana University   
School of Informatics and

Computing, Data Science  
Bloomington, IN

derrick@iu.edu

Michael Mzyk

Indiana University

School of Informatics and,

Computing, Data Science

Bloomington, IN

mmzyk@iu.edu

**ABSTRACT**

The Fortune 500 is the annual list of established companies, compiled and published by the Fortune magazine that ranks the 500 largest U.S. corporations by total revenue for their respective fiscal years. The Inc 5000 is the annual list of America’s fastest-growing private companies from fields like strategy, service and innovation, which is used in this study by its first 500 rankings. The aim of study is to analyze whether it is possible to predict if one company belonging to one of the former groups is “Fortune 500” or “Inc 500” by analyzing their Social Media behavior and usage data. For this purpose, social media usage data from Twitter and Facebook is collected and aggregated, and different descriptive and predictive models are built, evaluated and interpreted.

**Categories and Subject Descriptors**

G.3 [**Probability and Statistics**]: Statistical Computing; G.4 [**Mathematical Software**]: Algorithm Design and Analysis; H.3.3 [**Information Search and Retrieval**]: Search Process, Selection Process, Query Formulation, Information Filtering, Clustering; H.3.5 [**Online Information Services**]: Data Sharing, Web Based Services.

**General Terms**

Algorithms, Experimentation, Performance, Verification.

**Keywords**

Social Media, Social Media Mining, Twitter API, Facebook API, Fortune 500, Inc 500, Corporate Social Media

# INTRODUCTION

Our work is motivated by the research question as to whether there are significant similarities or differences in social media usage across different companies, and whether these patterns are predictable.

We focused on the most successful companies in the United States, belonging to the “Fortune 500” group of established businesses or to the “Inc 500” group of fast-growing businesses. We hypothesized that there were differences in their use of social media, due to cultural and historical differences and different levels of social media knowledge and expertise.

Our goal was to predict whether a company is a “Fortune 500” or “Inc 500” company based on its Facebook and Twitter behavior, by comparing the predictive power of Facebook and Twitter performance indicators.

# RELATED RESEARCH

We completed a review of the relevant literature, but only found little work that indicated whether there was a difference on how large, established businesses used social media as compared to younger, emerging businesses.

In her comparative study “How do the most successful companies use social media?” based on interviews of Chief Marketing Officers of leading companies, Ganim Barnes, N. [1] comes to the conclusion that while Social Media adoption in companies has increased constantly between 2007 and 2009, Inc. 500 companies have outpaced Fortune 500 companies in its use of social media such as Blogging, Tweeting and Social Networks (e.g. Facebook). Exactly the same findings could be reproduced four years later once again by Ganim Barnes, N., & Lescault, A. M. [2], highlighting for example that the pattern holds with 52% of the Inc. 500 companies blogging and 34% of the Fortune 500 companies. Although both studies represent first evidence that there are significant differences between both groups of companies regarding their social media usage, findings are not based on Social Media data itself (i.e., posts and tweets), and results are outdated.

In their research paper “We´re all connected: The Power of the Social Media Ecosystem”, Hanna, R., Rohm, A., & Crittenden, V.L. [3] point out that different social media platforms form an interconnected ecosystem rather than isolated websites, thus creating intimacy and engagement with customers actively influencing brand messages, products and services and serving as a “crystal ball” for predicting the future. They propose the tracking of key performance indicators (KPIs) indicating brand lift and brand engagement (such as “liking”) and combining them with “downstream metrics” such as sales wherever possible. Although the presented approach is more applied and hands-on, it does not differentiate between Fortune 500 and Inc. 500 companies and is not based on empirical studies.

In their research paper “Detecting and Analyzing Automated Activity on Twitter”, Zhang, C.M., & Paxson, V. [4] present a time-based prediction algorithm for detecting automated behaviour on Twitter, and identify that many accounts use a high degree of automation (e.g. by working with programmed schedules rather than being “spontaneous”). While we believe that automated scheduling of posts resp. tweets might be interesting variables when describing the social media activity of companies, the presented approach is rather general and does not refer to the social media behaviour of companies at all.

Finally, the research paper “Online Social Network Management Systems: State of The Art” written by Al-Qurishi, M., Al-Rakhami M., Al-Rubaian M., et al. [5] compares features of various social media management systems such as TweetDeck regarding their operational, security and online attributes and will be used as a starting point and guideline for our cross-platform approach.

# DATA COLLECTION AND PREPROCESSING

Company data describing all 1000 Fortune 500 and Inc 500 companies in scope of our analysis (cf. Chapter 3.1) was collected, and a maximum number of 125 Facebook posts (cf. Chapter 3.2) resp. Twitter tweets (cf. Chapter 3.3) have been pulled from the corresponding APIs together with related Social Media Indicators such as numbers of likes, comments etc. Based on this “raw” data, additional Features have been engineered (cf. Chapter 3.4), finally leading to three aggregated datasets to be further analyzed (Cf. Chapter 3.5).

## Company Data Collection

We first started with the collection of all Fortune 500 and Inc 500 company data from the year 2016. 2 datasets describing Fortune 500 companies and Inc. 500 companies were downloaded from public websites:

Table 1. The collected datasets

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Link for Download** | **Row**  **Count** |
| Fortune 500 | <https://knoema.com/FORTUNE500/fortune-500-companies> [6] | 500 |
| Inc 500 | <https://data.world/datanerd/inc-5000-2016-the-full-list>  (\*First 500 rows only) [7] | 500\* |

These datasets provide our main research entities (i.e., all 1000 companies “in scope”) and at the same time the “class labels” (i.e. “Facebook or Twitter”) of our final dataset, which will be predicted based on the Social Media performance indicators scraped from Facebook resp. Twitter.

Additionally, the company-specific features “Industry” (that a company belongs to) and “Revenue” (that was generated by the company) have been included in the dataset for being combined with Social Media KPIs for Cluster Analysis. For deriving “Industry” per company, 21 “sectors” (as found in the Fortune 500 dataset) and 24 “industries” (Inc. 500 dataset) have been summarized by only 8 new categories that are consistent with the “Global Industry Classification Standard.” [8] Those new categories can be seen in Appendix C.

## 3.2 Facebook Data Collection We performed the following extraction steps, with steps 3.- 5. automatically executed by a Python-based program establishing a connection to the Facebook API:

1. We manually identified the Facebook accounts for all 1,000 companies in scope. Automatically scraping them from company websites would not have been reliable, as some companies do not reference links to their Facebook accounts on their homepages.

2. We extracted Facebook addresses wherever necessary by clearing invalid characters, url paths, punctuations etc., thus obtaining a usable “Page Name” (e.g. “Vodafone” instead of “https://www.facebook.com/Vodafone&ref=ts”).

3. We confirmed all Facebook accounts and checked for existence and connectivity of webpages (some pages could not be reached by scrapers due to page-specific security settings).

4. We scraped all available posts for every identified and confirmed Facebook account from the Facebook API, with an upper limit of 125 posts per page, leading to a total number of 69923 scraped posts.

5. We scraped additional data for every post ID identified in former step from the corresponding pages, such as the number of likes or number of comments.

## 3.3 Twitter Data Collection

Similar to the Facebook process, we used a Python-based program for collecting the Twitter data by cronnecting to Twitter’s REST API for pulling out all relevant tweets.

1. We manually identified Twitter accounts for all 1,000 companies. Once again, scraping them from company websites would not have been reliable here, as some companies do not reference their Twitter links on their homepages.

2. Whenever necessary, we cleaned up Twitter addresses by remove extraneous characters that the API would not recognize (invalid characters, url paths, punctuations). In some occasions, this even included identifying the correct Twitter handle as some companies had a wrong or broken one on their homepage.

3. We requested the data via the Twitter REST API, including error checking and verifying that accounts were public and active. A handful of companies had official accounts that became private since then, and one company had its account suspended by Twitter for months for violating Twitter´s Terms of Service.

4. For every confirmed Twitter account, we requested the most recent 125 tweets in order to have an amount tweets comparable to the respective postings in the Facebook dataset, leading to a total number of 77148 tweets. We also pulled basic descriptor data from the tweet, such as text of tweet, number of likes, number of retweets etc.

5. We then requested biographic data about each twitter handle, including the number of followers that account had, as this would be important for determining amplification ability.

## 3.4 Feature Engineering

In addition to the basic “counts” (of friends, comments etc. on both company and posting level) that have been extracted so far, we thought of additional, derived ratio KPIs describing the relative engagement of the audience, inspired by the article “7 Social Media Metrics that Really Matter—and How to Track Them.” [9]

The 3 engagement KPIs (applicable to both Company and posting level) we additionally decided to implement for both our Facebook and Twitter datasets are:

* *Applause Rate*

This KPI can be defined as Avg. Ratio of likes per post/tweet by total number of likes.

In the article, the author suggested that instead of using stand-alone metrics such as “absolute” like counts, using the applause rate will give a better understanding for how much of the audiences for page likes the content.

* *Amplification Rate*

This feature can be defined as the average ratio of shares per post /retweets per tweet by total number of followers.

As described by the author, the amplification rate shows the willingness of the followers to associate with the brand, thus providing a clear picture of interest of an audience.

* *Conversation Rate*

This feature can be defined as the average ratio of comments per post/tweet divided by the total number of followers.

The author suggests this rate in order to determine the amount of audience that is contributed to the content. It is important for determining if the brand is really connected with the followers, having a conversation ongoing with it.

These three derived KPIs help to truly reflect the Social Media engagement of companies and to provide a more differentiated picture of Social Media usage.

## 3.5 Data aggregation

The workflow-based analytics tool KNIME [10] was used for aggregating the collected “raw” Company, Facebook and Twitter data (see chapters 3.1 – 3.3), performing all Feature Engineering steps (see chapter 3.4) and finally creating three comprehensive datasets that are described in the following.

As aggregation steps for deriving datasets on company level (i.e., with 500 rows for Fortune 500 and 500 rows for Inc 500) out of the raw data are straightforward, we refer the interested reader to Appendix A for documented screenshots of the complete KNIME workflow that has been implemented.

Additionally, we prepared the scraped datasets containing all postings for simple Sentiment Analysis by tagging positive or negative words per Facebook post resp. Twitter Tweet according to the MPQA Corpus [11]. Also, general word counts have been applied to all postings resp. tweets, and “Tag Clouds” have been implemented for visualizing results. When processing the data for Text Analysis, best practices of Text Analysis such as removal of common words, tokenization, conversion to uppercase, removal of punctuation marks and numbers, Usage of Bags of words etc. have been implemented. We refer again to Appendix A for documented screenshots of all text analysis workflows.

**3.5.1 Facebook Dataset**

The “Facebook Dataset” describes the Social Media behavior of both Fortune 500 and Inc. 500 companies on Facebook and contains KPIs defined on both the company level (e.g. “Number of likes per company”, describing how many likes the company Facebook homepage obtained) and post level (with aggregated KPIs over every post per company that has been scraped, usually by means of averaging results). This dataset will be used for predictive model building, answering the research question if a company is “Fortune 500” or “Inc. 500” based on its Facebook activity and behavior.

Table 2. Facebook KPIs

|  |  |
| --- | --- |
| **Column** | **Description** |
| COMPANY\_ID | ID of a company according to master file |
| FB\_POSTS\_NUM\_COMMENTS | The number of comments of a post |
| FB\_POSTS\_NUM\_LIKES | The number of likes of a post |
| FB\_POSTS\_NUM\_SHARES | The number of shares of a post |
| FB\_COMPANY\_NUM\_LIKES | Avg. Total number of likes per page, grouped into intervals |
| FB\_APPLAUSE\_RATE | Avg. Ratio of likes per post by total number of likes per page, grouped into intervals |
| FB\_AMPLIFICATION\_RATE\_PAGE | Avg. Ratio of shares per post by total number of likes per page, grouped into intervals |
| FB\_AMPLIFICATION\_RATE\_POST | Avg. Ratio of shares per post by total number of likes per post, grouped into intervals |
| FB\_CONVERSATION\_RATE\_PAGE | Avg. Ratio of comments per post to the number of Page Likes, grouped into intervals |
| FB\_CONVERSATION\_RATE\_POST | Avg. Ratio of comments per post to the number of Post Likes, grouped into intervals |
| Dataset | Class label with values Fortune 500 or Inc. 500 |

**3.5.2 Twitter Dataset**

The “Twitter Dataset” describes the Social Media behavior of both Fortune 500 and Inc. 500 companies on Twitter and contains KPIs defined on both the company level (e.g. “Number of followers per company”) and tweet level (with aggregated KPIs over every tweet per company that has been scraped, usually by means of averaging results). This dataset has almost the same KPIs as the Facebook Dataset, making the behavior of a company on Facebook and Twitter quite comparable. This dataset will be used for predictive model building as well, answering the research question if a company is “Fortune 500” or “Inc. 500” based on its Twitter activity and behavior.

Table 3. Twitter KPIs

|  |  |
| --- | --- |
| **Column** | **Description** |
| COMPANY\_ID | ID of a company according to master file |
| TW\_COMPANY\_FRIENDS | The number of other accounts that a company is following |
| TW\_COMPANY\_FOLLOWERS | The number of followers of the company |
| TW\_TWEETS\_NUM\_REPLIES | Avg. Total number of replies per tweet grouped into intervals |
| TW\_TWEETS\_NUM\_FAVORITES | Avg. Total number of favorites per tweet grouped into intervals |
| TW\_TWEETS\_NUM\_RETWEETS | Avg. Total number of retweets per tweet grouped into intervals |
| TW\_APPLAUSE\_RATE | Avg. Ratio of likes per tweet by total number of followers, grouped into intervals |
| TW\_AMPLIFICATION\_RATE\_PAGE | Avg. Ratio of retweets per tweet by total number of followers, grouped into intervals |
| TW\_AMPLIFICATION\_RATE\_POST | Avg. Ratio of retweets per tweet by total number of likes per tweet, grouped into intervals |
| TW\_CONVERSATION\_RATE\_PAGE | Avg. Ratio of comments per tweet to the number of followers, grouped into intervals |
| TW\_CONVERSATION\_RATE\_POST | Avg. Ratio of comments per tweet to the number of Tweet Likes, grouped into intervals |
| Dataset | Class label with values Fortune 500 or Inc. 500 |

**3.5.3 Company Characteristics Dataset**

The “Company Characteristics Dataset” contains further features that characterizes a company (Revenue per Employee and Industry), as well as two Flags indicating if the company has a Facebook or Twitter account, respectively. It also contains the “Dataset” feature indicating if a company belongs to the “Fortune 500” or “Inc. 500” group. However, this feature does not represent a class label in this dataset but rather one more, “ordinary” feature, as the Dataset will be used for cluster analysis exclusively.

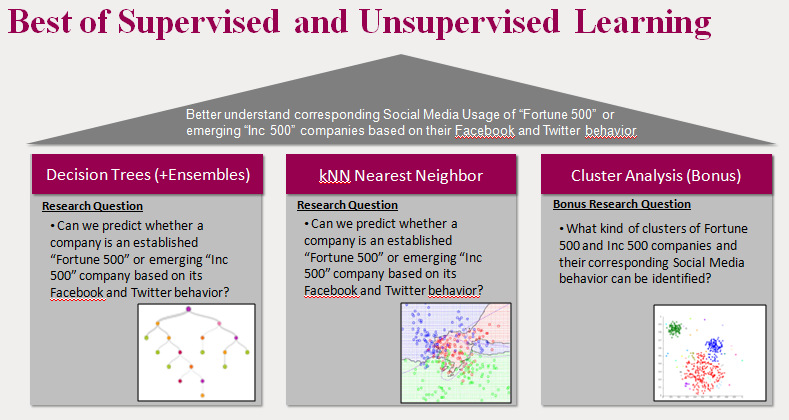
Table 4. Company Features

|  |  |
| --- | --- |
| **Column** | **Description** |
| COMPANY\_ID | ID of a company according to master file |
| REVENUE\_PER\_EMPLOYEE | Revenue of company, divided by number of employees |
| INDUSTRY | Vertical/Industry of company (e.g. “Energy”, “Health”) |
| FB\_ACCOUNT\_FLAG | Indicator if company has an active Facebook account |
| TW\_ACCOUNT\_FLAG | Indicator if company has an active Twitter account |
| Dataset | Indicator if company is Fortune 500 or Inc. 500 |

# ANALYTICAL MODELS

A combination of different descriptive and predictive models (covering both Supervised and Unsupervised Learning algorithms) has been applied to the three datasets for the purpose of a multi-facetted and complete picture of different dimensions of Social Media Usage of Fortune 500 vs. Inc. 500 companies (see Figure 1).

Figure 1. Overview of Analytical Models applied



All analytical models have been build and tuned based on the Open Source Data Mining Software Weka [12] from Waikato University.

Additionally, simple word counts and Sentiment Analysis have been performed for getting a better idea of data semantics and usage of emotions along Fortune 500 vs. Inc. 500 companies.

## Text and Sentiment Analysis

As a starting point, simple Sentiment Analysis was performed by tagging positive or negative words per Facebook post resp. Twitter tweet according to the MPQA Corpus [11]. Based on individual word counts of these words per post resp. tweet, global word counts were derived and finally visualized as Tag Clouds representing the most frequent positive/negative terms, with one Tag Cloud for “Fortune 500” words and another one for “Inc. 500” words exclusively. Additionally, Tag Clouds representing all terms were be provided in order to obtain a broader picture of relevant posting contents.

Note that all “Tag Clouds” used for visualizing those results were already part of the KNIME workflow (see Chapter 3.5).

## Simple Classifiers

As a next step, we built simple classifiers in order to analyze our main research question, i.e. whether it is possible to predict whether a company is an established “Fortune 500” company versus an emerging “Inc. 500” company, based on its Social Media behavior.

All classifiers were applied to both the Facebook dataset (see Chapter 3.5.1) and the Twitter dataset (see Chapter 3.5.2). Taking into consideration the general characteristics of these datasets (only numeric features, and binary class label), we considered the following Weka implementations [12] of simple classifiers for our prediction task:

* J48 Decision Tree with minnumobj = 2 and confidenceFactors 0.25 (standard pruning) and 0.01 (lots of pruning)
* kNN Nearest Neighbor with k = 1 and k = 5
* Logistic Regression

Due to the small amount of rows in the dataset (1000), all algorithms were run using 10 fold cross validation, while avoiding a split of the datasets into Training and Test data.

Note that the following features of the Facebook and Twitter datasets were explicitly excluded from analysis before running the algorithms:

* COMPANY\_ID (in both datasets)
* FB\_COMPANY\_NUM\_LIKES
* TW\_COMPANY\_FOLLOWERS

While the reasons for excluding the (“random”) ID column “COMPANY\_ID” were quite obvious, the company-level KPIs FB\_COMPANY\_NUM\_LIKES (from Facebook dataset) and TW\_COMPANY\_FOLLOWERS (from Twitter dataset) were taken out, as - just “by themselves” - they already were highly accurate predictors for the “Fortune 500 vs Inc 500” research question with predictive accuracies of around 85%, outperforming all other features in terms of predictive power and leading to little attractive / diversified results (Fortune 500 companies simply have a much larger base of followers than Inc. 500 companies in absolute terms, as they are far better known and established).

Leaving FB\_COMPANY\_NUM\_LIKES and TW\_COMPANY \_FOLLOWERS out of analysis enabled us to better understand the relative importance of all other KPIs (both absolute counts on post level such as “Number of shares” and relative ratios such as “Applause Rate”), especially in case of “white box” algorithms such as Decision Trees that lead to a set of actionable rules.

**4.3 Ensemble Methods**

In addition to the simple classifiers, the following ensembles of trees have been implemented applied to the datasets (under the same configurations and restrictions described in the former chapter):

* Random Forests
* Decision Tree Bagging for all J48 Decision Trees as described above

Ensemble Methods were used under the hypothesis of potentially improving accuracies and ROC values of the simple classifiers.

# RESULTS

## Text and Sentiment Analysis

We found out that Inc 500 companies use words in their Social Media postings more regularly that are associated with emotional extremes, such as “happy”, “great”, “excited” or “amazing”. In case of Fortune 500 companies, while also using words like “happy” and “great”, to a lesser extent, the remaining words are more conservative in nature (e.g. “support”, “proud”, “commitment”, “sustainable”).

Note that these findings are consistent for both “Facebook” (see Fig. 2) and “Twitter” platforms, which have been analyzed separately.

Figure 2. Fortune 500 vs. Inc. 500 Sentiment Analysis (FB)



As a “rule of thumb”, if a posting conjures a strong emotion, it is probably from an Inc 500 company.

While being less characteristic and clear, general trends in word counts (i.e. of all words, not only of positive or negative ones) hint at different collaboration models (Fort500: “World”, “Employee” vs. Inc500: “Market”, “Service”) and planning horizons (Fort500: “Year”, Inc500: “Day”)

Once again, these findings are consistent for both “Facebook” (see Fig. 3) and “Twitter” platforms, which have been analyzed separately.

Figure 3. Fortune 500 vs. Inc. 500 General Word Counts (FB)

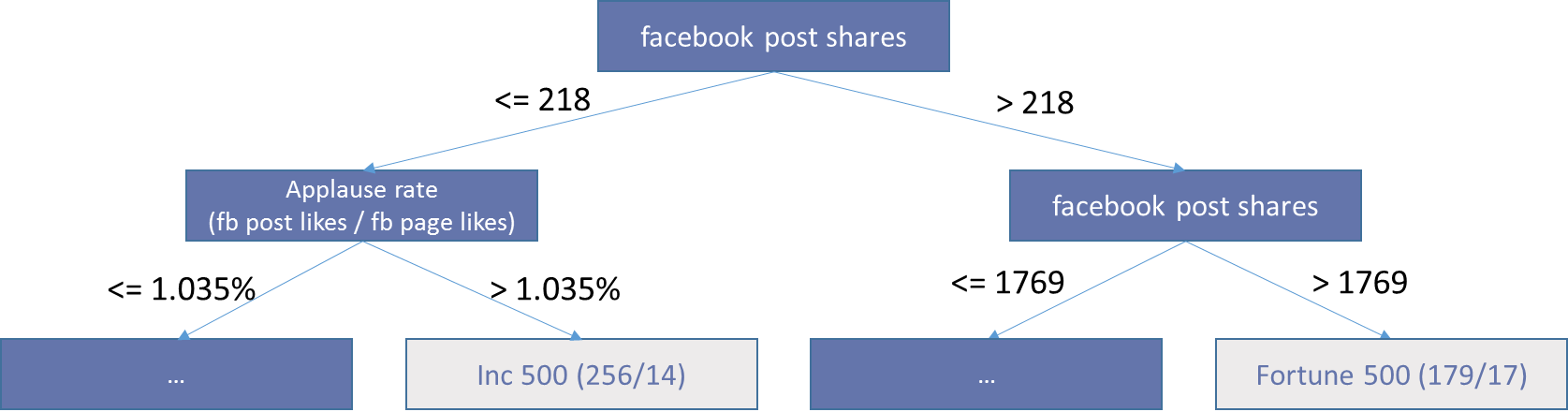


Please refer to Apendix X for a complete overview of all Tag Clouds (including those for the Twitter platform) that could be derived.

## Simple Classifiers

Most simple classifiers achieved high accuracies of around 80% and good ROC values (with ROC referring to the area under the ROC curve). For both the Facebook and Twitter datasets, J48 decision trees (cf. Appendix D) with a standard confidence factor of 0.25 in Weka performed the best, achieving 84.3% accuracy (and 0.854 ROC) for Facebook accounts versus 86.7% accuracy (and 0.866 ROC) for Twitter.

Figure 4. J48 Decision Tree (FB), first 2 levels



For Facebook, the top-level node was the number of shares. The more shares, the increased likelihood it was a Fortune 500 company. Similarly, for Twitter, a high number of likes was indicative for Fortune 500 companies. These is a result of Fortune 500 companies tending to have much larger audiences.

The decision trees also show where Inc 500 companies are able to distinguish themselves. They tended to perform better with our derived features such as applause rate and amplification rate. These numbers are based on per capita likes (Facebook page) or followers (Twitter). While having smaller audiences, Inc 500 sites resonate better and have more engagement with their followers, which on average are more active and participative.

Logistic Regression provided classification accuracies of 78% for Facebook pages and 81.4% for Twitter accounts. While this result is still quite significant, it was not the most predictive method. Moreover, results of this algorithm were less interpretable, as it is a “black box algorithm”, whereas in case of J48, a concrete set of rules (resp. “tree”) could be derived, thus directly providing actionable advise.

In a similar fashion, the k-Nearest Neighbors algorithm (with k=5) provided classification accuracies of 72% for Facebook and 77% for Twitter.

Complete results of all simple classifiers under different tuning parameters can be viewed in Appendix E.

## Ensemble Methods

Not surprisingly, the highest accuracies and ROC values measured in all predictive models came from the ensemble methods as additional improvement of “simple” trees. The classification accuracy of Random Forests applied to Facebook pages was 86% (with ROC 0.921), while for Twitter accounts it was at 88.5% (with ROC 0.949). However, just as in case of Logistic Regression and k-means clustering, the generated output is less interpretable, as we are dealing with a whole “wood” of random trees instead of just a single one.

The Decision Tree Bagging algorithm (applied to our “best” J48 Decision Trees, i.e. those with standard confidence factor of 0.25) also led to excellent performance results for Facebook (Accuracy 84.9%, ROC 0.915) and Twitter (Accuracy 87.6%, ROC 0.93). Performance results were just slightly below those for Random Forests, possibly due to the lack of the random component.

Complete results of all the ensemble methods classifiers can be viewed in Appendix E.

# CONCLUSIONS

By all accounts, it is clear that through their behavior on social media sites, Fortune 500 and Inc 500 companies are predictably classifiable.

Social media postings are yet another field where well-established companies (Fortune 500) are predictably conservative, and fast-growing companies (Inc 500) take more risks. This study provides evidence that there are significant, predictable differences on how different types of companies use social media.

We found that Inc 500 companies use words in their social media posting more regularly that are associated with emotional extremes. Inc 500 companies are rewarded for their risky behavior, with higher engagement rates with their followers, resulting in a much more interactive experience with their customers.

On the other hand, Fortune 500 companies do leverage their history and brand recognition in order to reach much broader audiences that Inc 500 companies can only dream about it.

In conclusion, if a posting conjures a strong emotion, it’s probably from an Inc 500 company. If a lot have people seen it, it likely originated from a Fortune 500 company.

We observed that some companies had invested in other social media platforms (LinkedIn, YouTube, Snapchat, blogging, etc) Future research should see if inclusion of these could can potentially improve results.

We selected the most recent 125 postings in each social media platform. Some companies hit this within a day or two, and others took years. Future research should determine if using a time-based approach instead of a postings-based approach could improve results.

In addition, given the robust data

Lastly, a cross-domain analysis of Facebook and Twitter (and/or other social media sites) could improve results.

# TEAM MEMBER CONTRIBUTIONS

Aydan Alyanak pulled the Fortune 500 biographic data, preprocessed the social media handles of the companies, performed the API calls from the Facebook API for all 1000 companies, and jointly lead the final report with Derrick.

Derrick Eckardt pulled the Inc 500 biographic data, performed the API calls from the Twitter API for all 1000 companies, lead the drafting of the project poster, and jointly lead the final report with Aydan.

Michael Mzyk drafted the projected proposal and the project report, preprocessed the data using KNIME, performed descriptive and predictive analysis in Weka, and drafted the corresponding sections in the poster and final report.

# ACKNOWLEDGMENTS

Our thanks go to Vincent Malic and Ashley Dainas for their support and feedback as we progressed through the project.

# REFERENCES

|  |  |
| --- | --- |
| [1] | M. Ganim Barnes, "How do the most successful companies use social media?," *Marketing Research,* pp. 8-13, Spring 2010. |
| [2] | N. &. L. A. M. Ganim Barnes, "The 2013 Inc. 500 and Social Media.," University of Massachusetts Dartmouth Center for Marketing Research, 2014. |
| [3] | R. R. A. &. C. V. Hanna, "We´re all connected: The Power of the Social Media Ecosystem," *Business Horizons,* vol. 54, pp. 265-273, 2011. |
| [4] | C. &. P. V. Zhang, "Detecting and Analyzing Automated Activity on Twitter," in *PAM'11 Proceedings of the 12th international conference on Passive and active measurement*, 2011. |
| [5] | M. A.-R. ,. M. A.-R. M. e. a. Al-Qurishi, "Online Social Network Management Systems: State of The Art," in *Proceedings of the International Conference on Advanced Wireless Information and Communication Technologies 2015 (AWICT)*, 2015. |
| [6] | Fortune, "Fortune 500 Companies," Knoema.com, 31 May 2916. [Online]. Available: https://knoema.com/FORTUNE500/fortune-500-companies. [Accessed 15 February 2017]. |
| [7] | Inc Magazine, "Inc 5000 2016: The Full List - dataset by datanerd," data.world, 30 August 2016. [Online]. Available: https://data.world/datanerd/inc-5000-2016-the-full-list. [Accessed 15 February 2017]. |
| [8] | MSCI, "Global Industry Classification Standard (GICS®)," MSCI, 2016. |
| [9] | S. Dawley, "7 Social Media Metrics that Really Matter—and How to Track Them," 29 July 2016. [Online]. Available: https://blog.hootsuite.com/social-media-metrics/. [Accessed 1 March 2017]. |
| [10] | KNIME, "KNIME | Open for Innovation," [Online]. Available: https://www.knime.org/. [Accessed 15 April 2017]. |
| [11] | MITRE Corporation, "MPQA Opinion Corpus | MPQA," University of Pittsburgh, [Online]. Available: http://mpqa.cs.pitt.edu/corpora/mpqa\_corpus/. [Accessed 15 April 2017]. |
| [12] | Machine Learning Group at the University of Waikato, "Weka 3 - Data Mining with Open Source Machine Learning Software in Java," [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 15 April 2017]. |

# APPENDIX A – KNIME WORKFLOWS

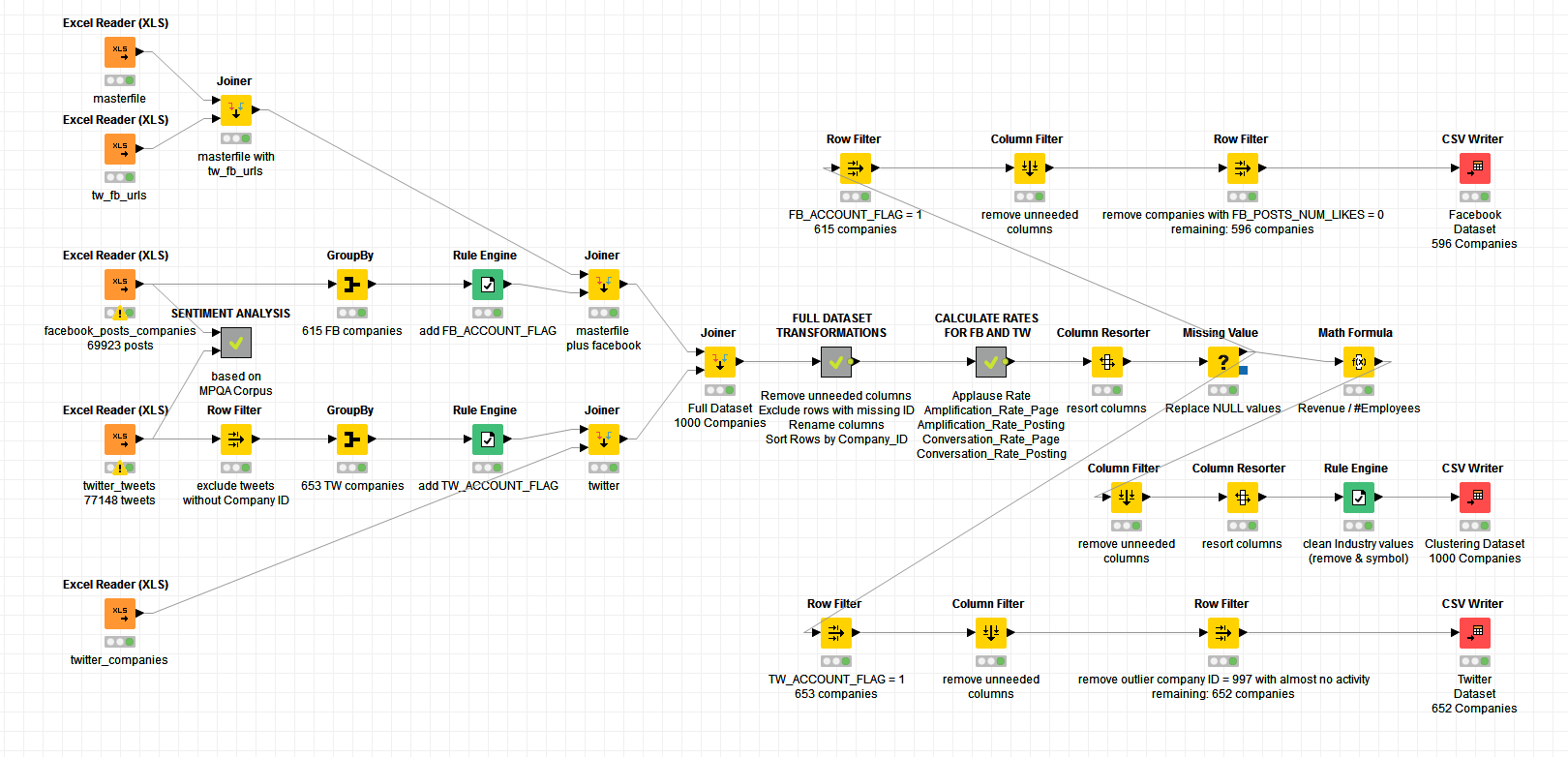


Figure A1. Main KNIME workflow utilized during this report.

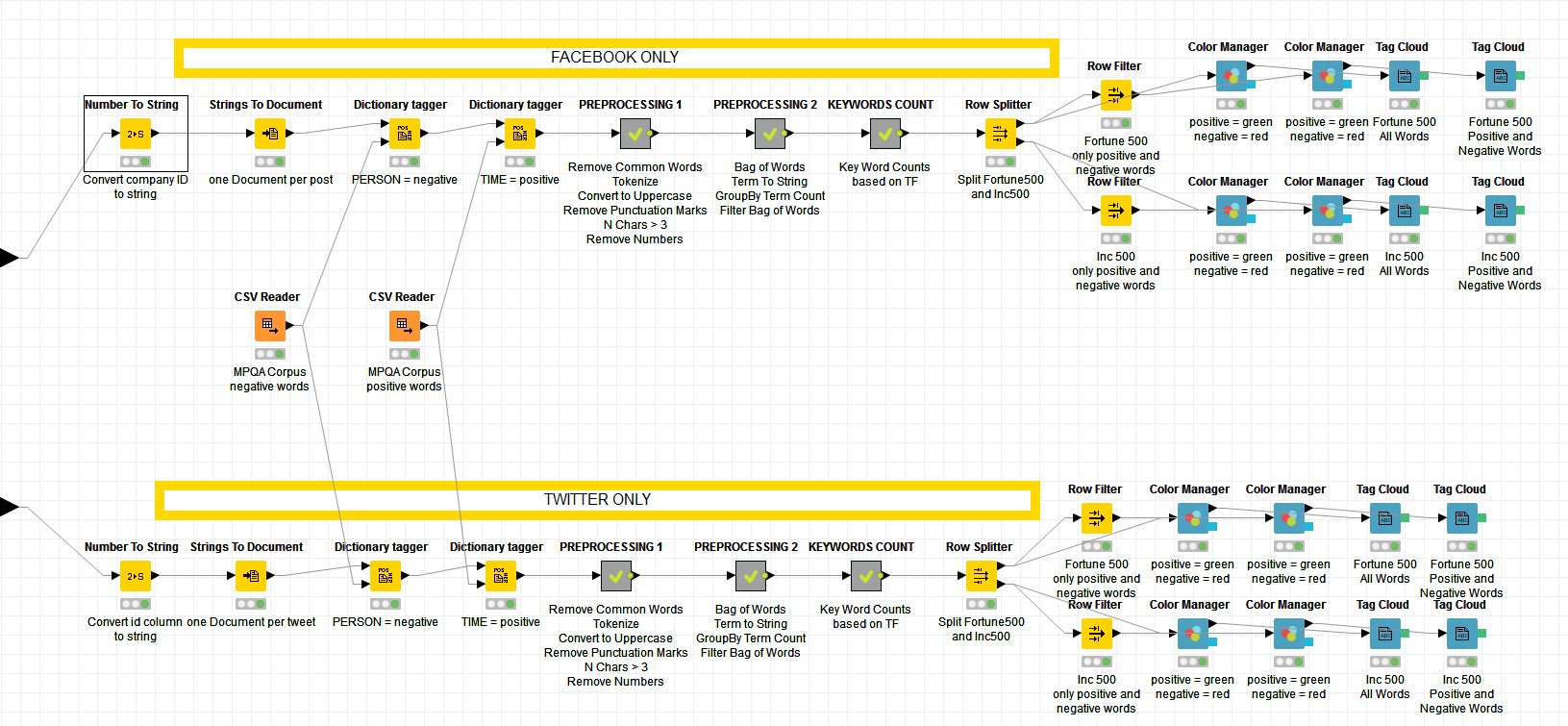


Figure A2. Metanode Sentiment Analysis

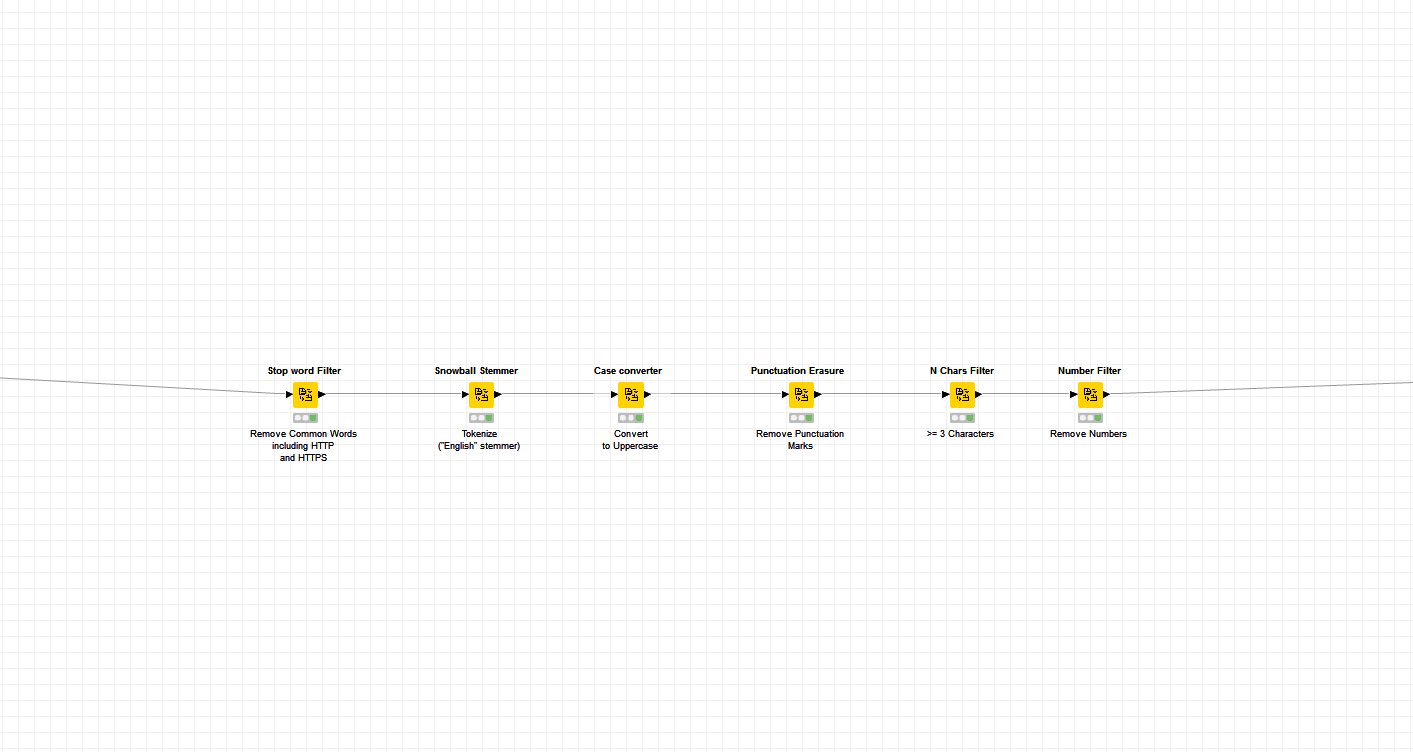


Figure A3. Metanode Preprocessing, Part 1

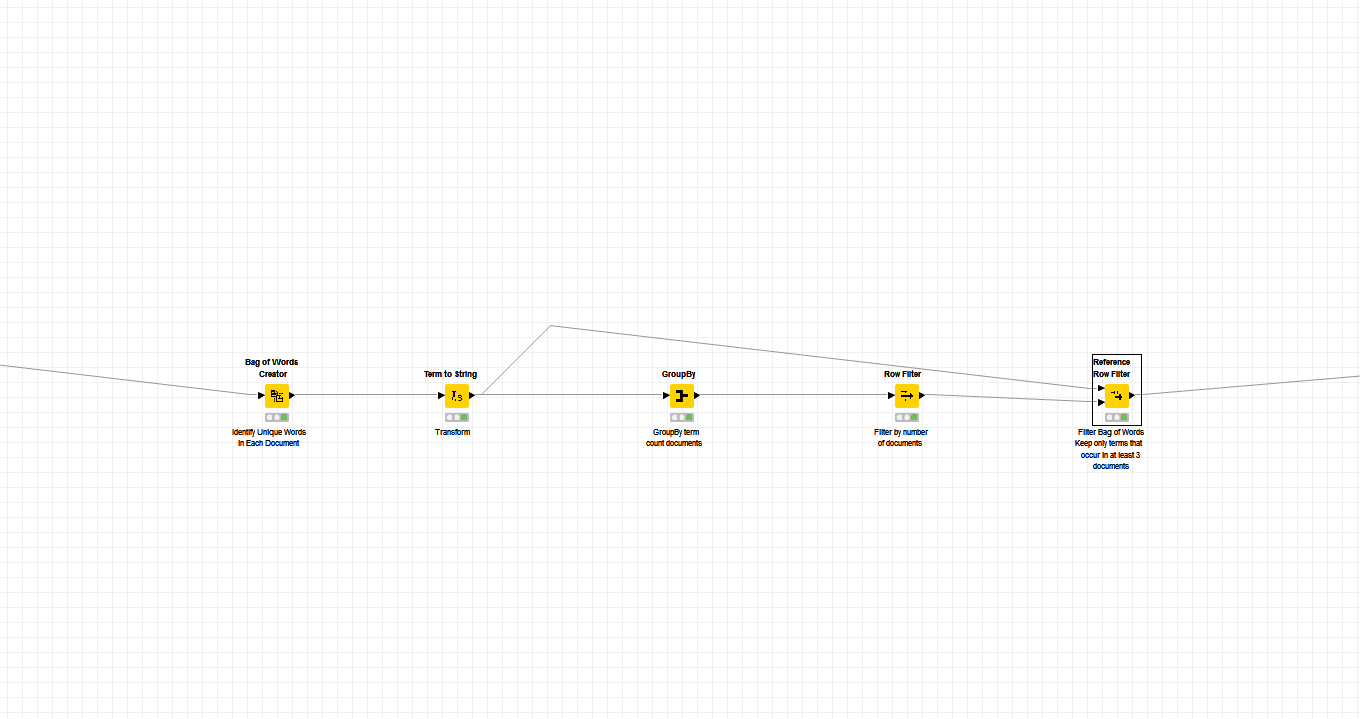


Figure A4. Metanode Preprocessing, Part 2

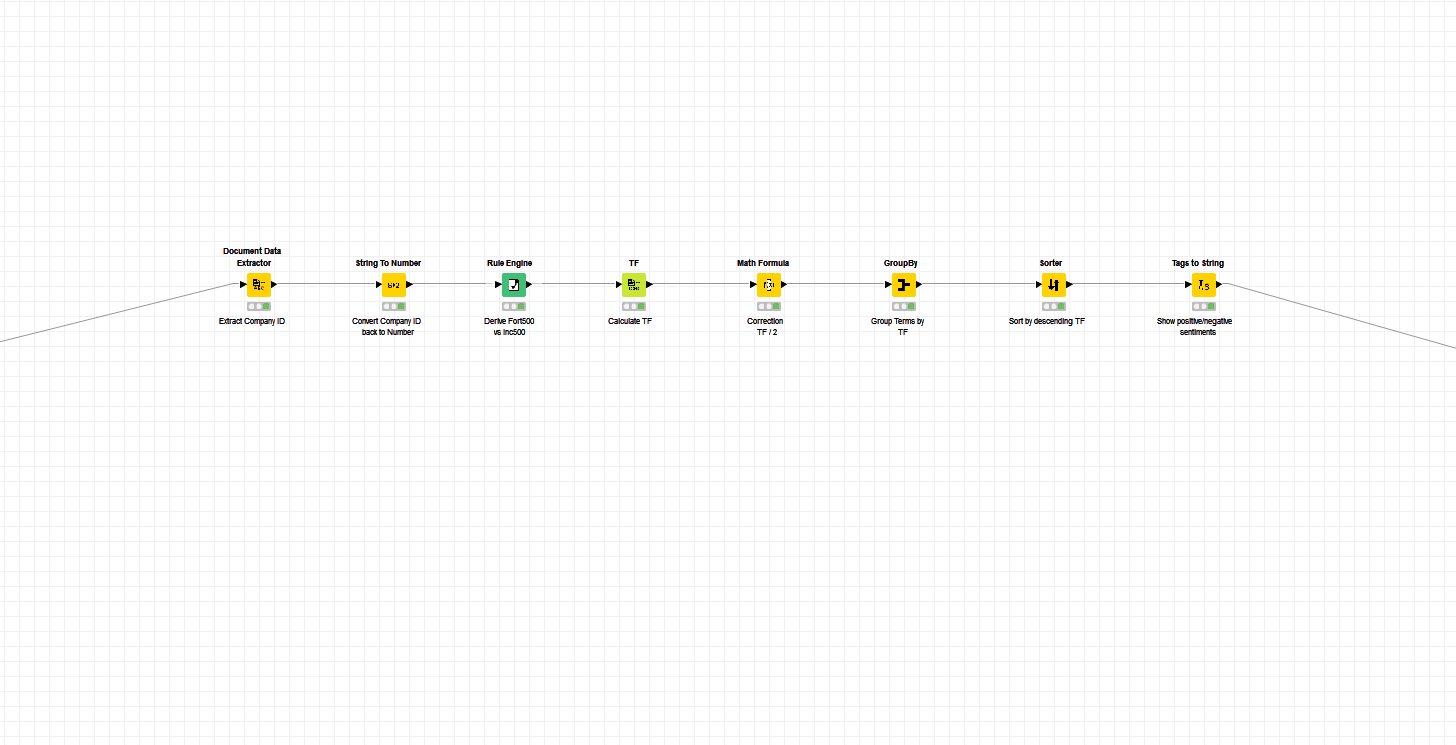


Figure A5. Metanode Keyword Count

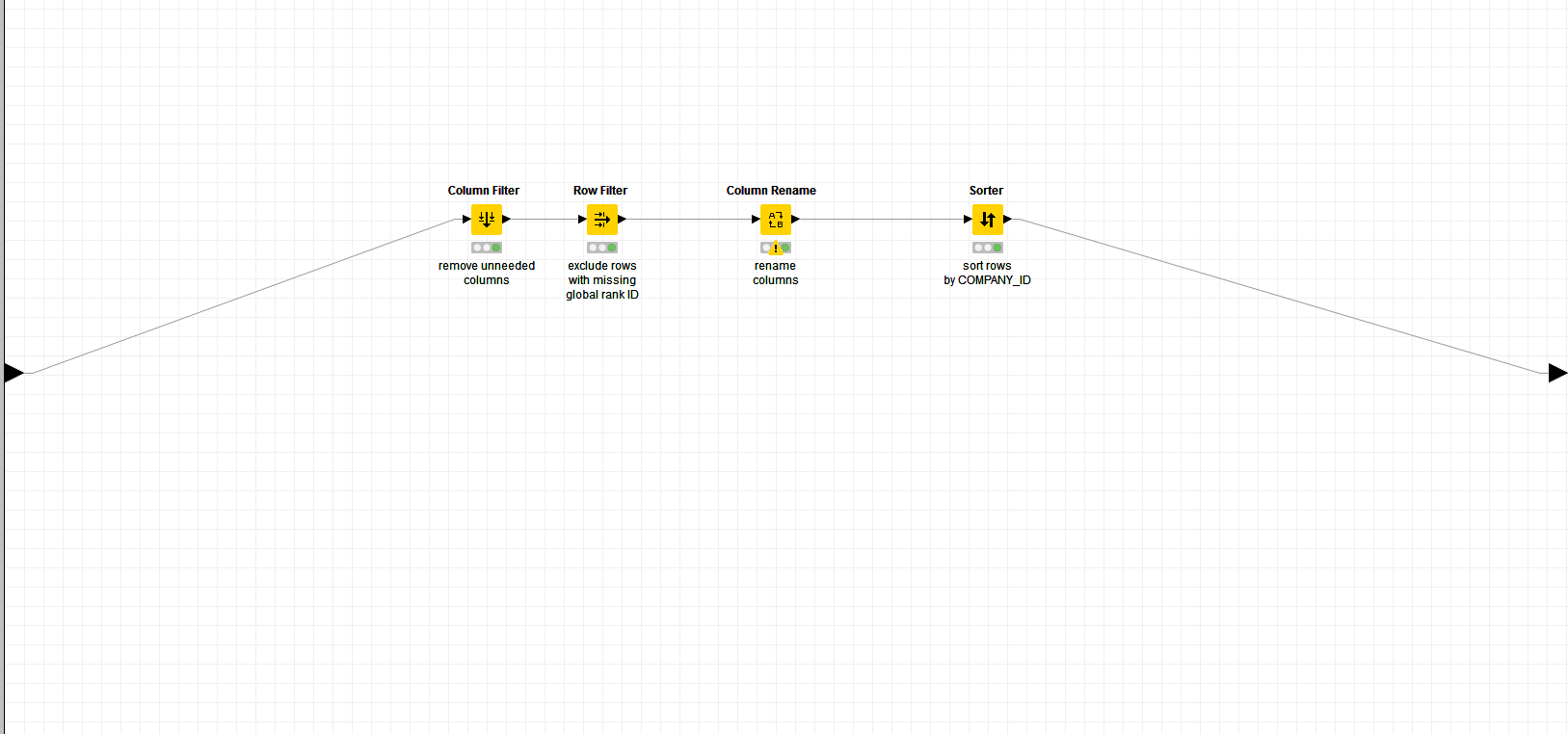


Figure A6. Metanode Full Dataset Transformations

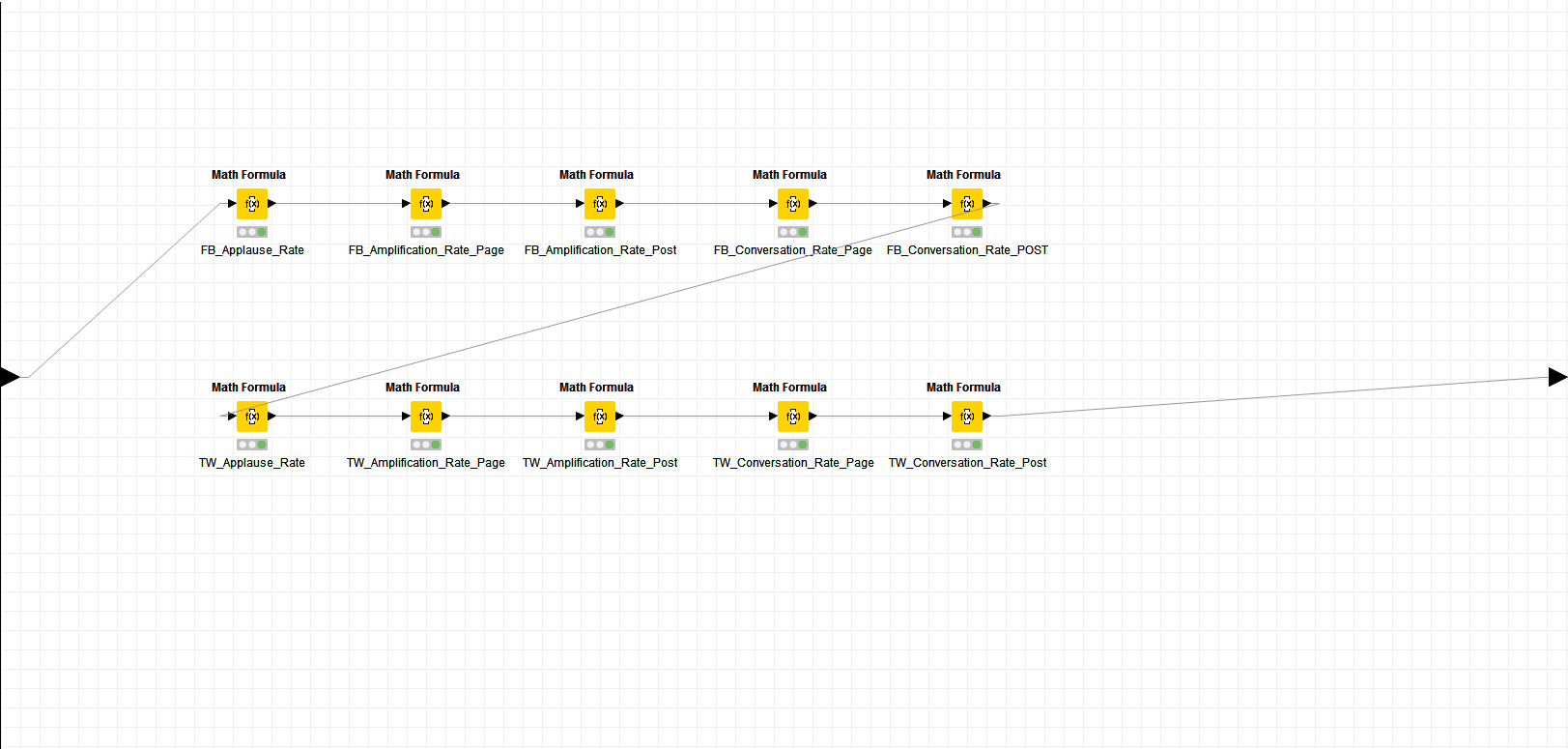


Figure A7. Metanode Calculate Rates for Facebook and Twitter

# APPENDIX B – SENTIMENT ANALYSIS WORD CLOUDS

|  |  |
| --- | --- |
|  |  |
| Figure B1. Tag Cloud, Facebook, Fortune 500, All Words | Figure B2. Tag Cloud, Facebook, Inc 500, All Words |
|  |  |
| Figure B3. Tag Cloud, Facebook, Fortune 500, Positive and Negative Words | Figure B4. Tag Cloud, Facebook, Inc 500, Positive and Negative Words |

|  |  |
| --- | --- |
|  |  |
| Figure B5. Tag Cloud, Twitter, Fortune 500, All Words | Figure B6. Tag Cloud, Twitter, Inc 500, All Words |
|  |  |
| Figure B7. Tag Cloud, Twitter, Fortune 500, Positive and Negative Words | Figure B8. Tag Cloud, Twitter, Inc 500, Positive and Negative Words |

# APPENDIX C – INDUSTRY COMPARISON OF FORTUNE 500 and INC 500

Table C1. – Industry Comparison of Fortune 500 and Inc 500.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Industry (Fortune 500, Inc 500)** | **#** | **Sector (Fortune 500)** | **#** | **Industry (Inc 500)** | **#** |
| Energy | 104 | Energy | 88 | Energy | 16 |
| Industrials & Materials | 186 | Transportation | 21 | Logistics & Transportation | 14 |
|  | Engineering & Construction | 13 | Engineering | 7 |
|  |  |  | Construction | 21 |
|  | Motor Vehicles & Parts | 34 | Manufacturing | 7 |
|  | Aerospace & Defense | 15 | Security | 7 |
|  | Industrials | 21 |  |  |
|  | Materials | 18 |  |  |
|  | Chemicals | 8 |  |  |
| Business and Government Products & Services | 74 | Business Services | 1 | Business Products & Services | 39 |
|  |  |  | Government Services | 27 |
|  |  |  | Human Resources | 7 |
| Consumer Products & Services | 169 | Food & Drug Stores | 20 | Consumer Products & Services | 29 |
|  | Food, Beverages & Tobacco | 17 | Food & Beverage | 28 |
|  | Household Products | 2 | Education | 10 |
|  | Hotels, Restaurants & Leisure | 5 | Travel & Hospitality | 6 |
|  | Media | 3 | Media | 7 |
|  |  |  | Advertising & Marketing | 42 |
| Retail & Wholesale | 72 | Retailing | 18 | Retail | 28 |
|  | Wholesalers | 23 |  |  |
|  | Apparel | 3 |  |  |
| Health | 81 | Health Care | 27 | Health | 54 |
| Financials, Insurance & Real Estate | 171 | Financials | 113 | Financial Services | 36 |
|  |  |  | Insurance | 7 |
|  |  |  | Real Estate | 15 |
| Technology & Telecommunications | 143 | Technology | 33 | Computer Hardware | 3 |
|  |  |  | IT Services | 47 |
|  |  |  | Software | 40 |
|  | Telecommunications | 17 | Telecommunications | 3 |
| TOTAL | 1000 |  | 500 |  | 500 |

# APPENDIX D – DECISION TREES

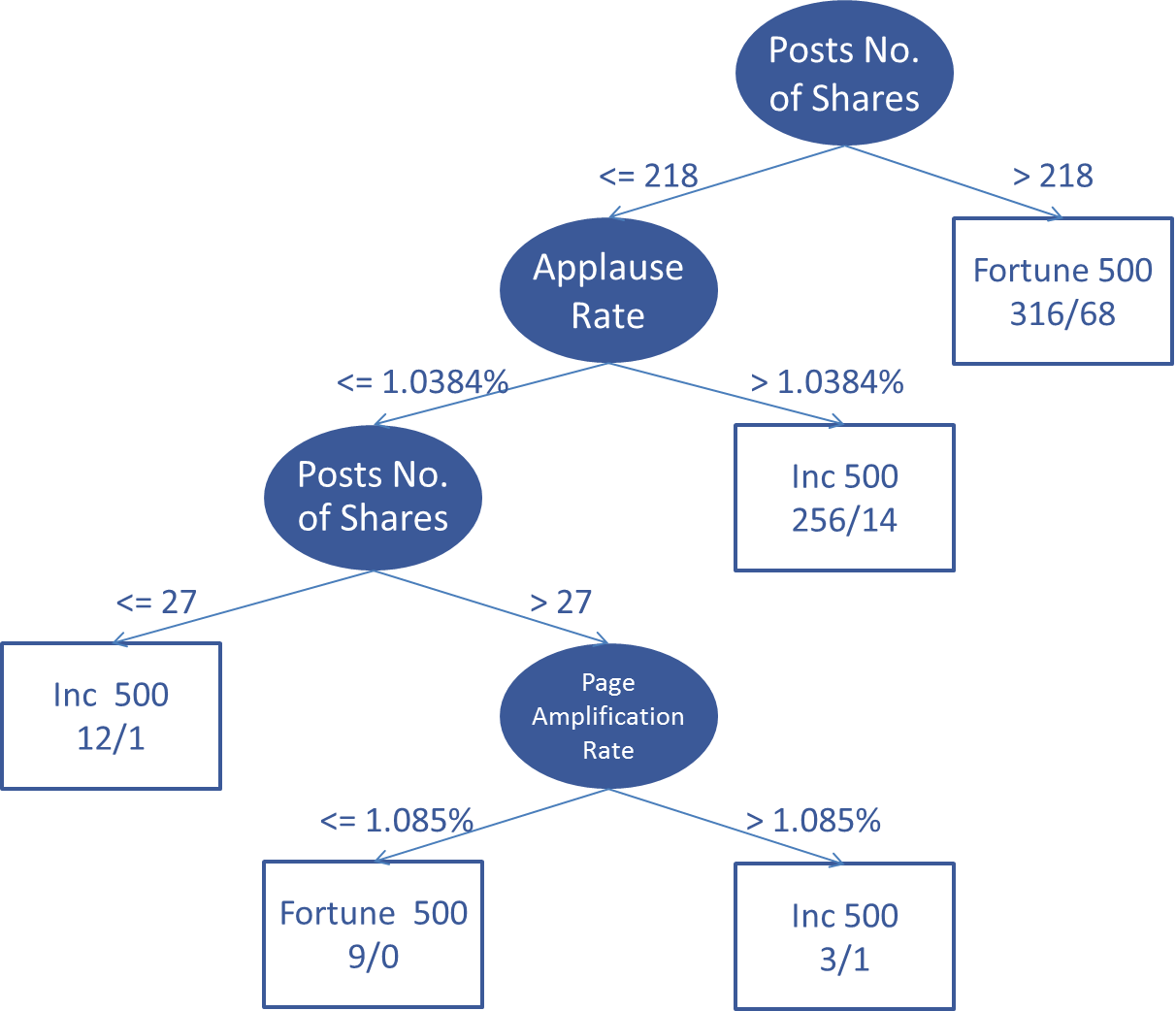


Figure D1. Final Facebook Decision Tree

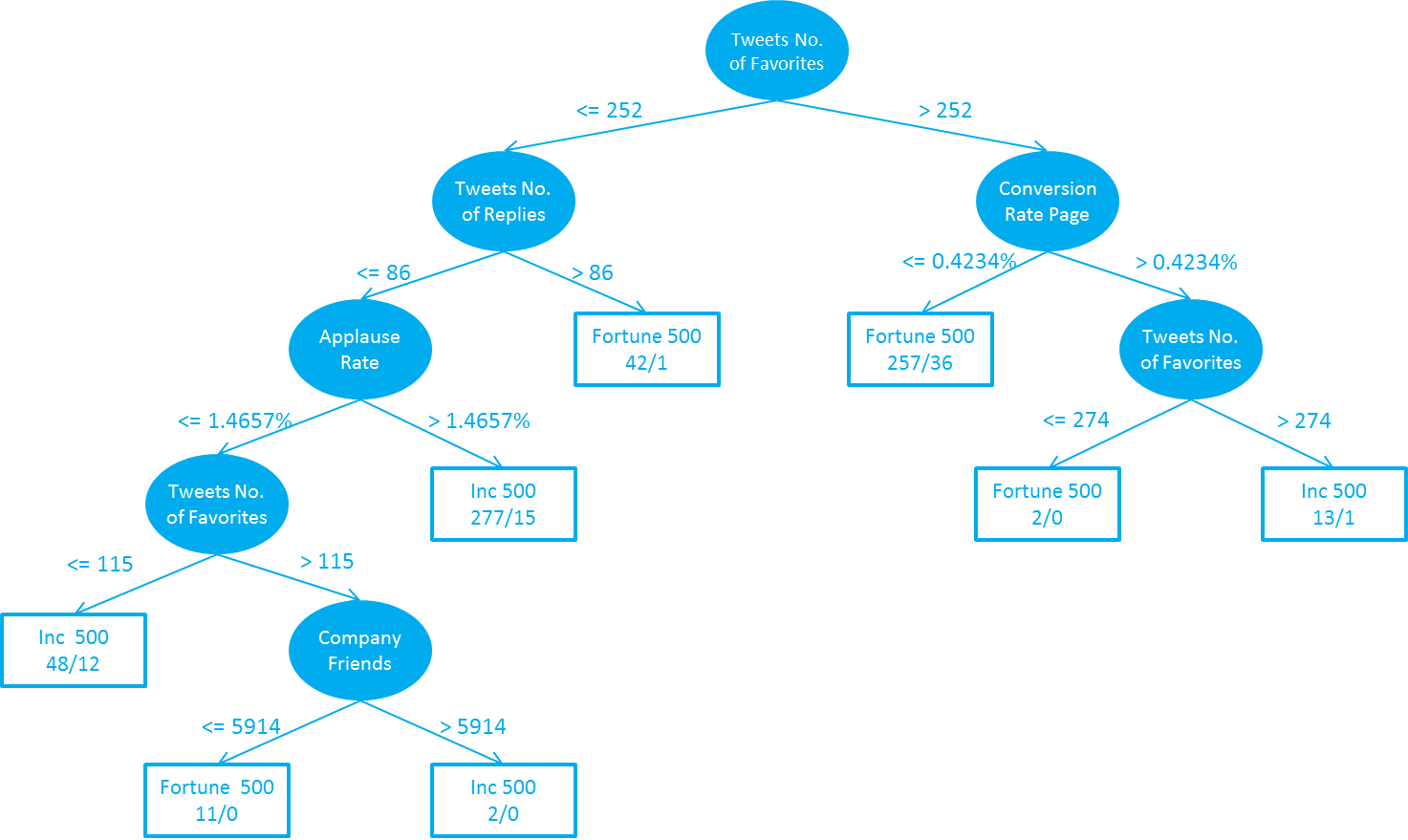


Figure D2. Final Twitter Decision Tree

# APPENDIX E – SUMMARY TABLE OF SIMPLE AND ENSEMBLE CLASSIFIERS

Table E1. – Summary of Simple and Ensemble Classifiers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Algorithm** | **Accuracy** | **ROC Area** | **Interpretability** |
| Facebook Dataset  10-Folds Cross Validation  Features excluded from dataset:  COMPANY\_ID  FB\_COMPANY\_NUM\_LIKES | J48 Decision Tree,  confidenceFactor = 0.25, minnumobj = 2 | 84.2% | 0.854 | high |
| J48 Decision Tree  confidenceFactor = 0.01, minnumobj = 2 | 83.6% | 0.834 | high |
| kNN with k = 1 | 72.1% | 0.725 | low |
| kNN with k = 1 and normalized data | 72.1% | 0.725 | low |
| kNN with k = 5 | 72.0% | 0.752 | low |
| kNN with k = 5 and normalized data | 72.0% | 0.752 | low |
| Logistic | 78.0% | 0.877 | low |
| Logistic and normalized data | 78.0% | 0.877 | low |
| DT Bagging J48  confidenceFactor = 0.25, minnumobj = 2 | 84.6% | 0.915 | low |
| DT Bagging J48  confidenceFactor = 0.01, minnumobj = 2 | 84.9% | 0.903 | low |
| RandomForest | 86.0% | 0.921 | low |
| Twitter Dataset  10-Folds Cross Validation  Features excluded from dataset: Company\_ID  TW\_COMPANY\_FOLLOWERS | J48 Decision Tree  confidenceFactor = 0.25, minnumobj = 2 | 86.7% | 0.866 | high |
| J48 Decision Tree  confidenceFactor = 0.01, minnumobj = 2 | 86.5% | 0.873 | high |
| kNN with k = 1 | 76.0% | 0.762 | low |
| kNN with k = 1 and normalized data | 76.0% | 0.762 | low |
| kNN with k = 5 | 77.1% | 0.849 | low |
| kNN with k = 5 and normalized data | 77.1% | 0.849 | low |
| Logistic | 81.4% | 0.884 | low |
| Logistic and normalized data | 81.4% | 0.884 | low |
| DT Bagging J48, confidenceFactor = 0.25, minnumobj = 2 | 87.6% | 0.93 | low |
| DT Bagging J48, confidenceFactor = 0.01, minnumobj = 2 | 87.3% | 0.924 | low |
| RandomForest | 88.5% | 0.949 | low |

# APPENDIX F – SENTIMENT ANALYSIS WORD CLOUDS